DOCUMENT RESUME

ED 382 509 SO 024 930

TITLE	Setting Achievement Levels on the 1994 National Assessment of Educational Progress in Geography and in U.S. History and the 1996 National Assessment of Educational Progress in Science. Final Version. Design Document.
INSTITUTION	American Coll. Testing Program, Iowa City, Iowa.
SPONS AGENCY	National Assessment Governing Board, Washington, DC.
PUB DATE	Apr 94
CONTRACT	ZA9003001
NOTE	98p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Academic Achievement; *Educational Assessment; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; Evaluation Research; Geography; Measurement; Minimum Competencies; Science Instruction; Standards; Student Evaluation; United States History
IDENTIFIERS	*National Assessment of Educational Progress

ABSTRACT

This design document describes a research methodology for setting achievement levels (performance standards) on the NAEP (National Assessment of Educational Progress). The current plan is an extension of earlier work by NAGB (National Assessment Governing Board) to set achievement levels on the 1990 NAEP mathematics assessment, and by NAGB and American College Testing (ACT) to set achievement levels on the 1992 NAEP assessments in mathematics, reading, and writing. The volume contains the five sections: (1) Identification and Selection of Panelists; (2) Preparation of Briefing Materials; (3) Achievement Levels)Setting Procedures; (4) Statistical Analyses; and (5) Public Comment Forums. A list of references and appendices are included. (EH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ACT

ED 382 509

DESIGN DOCUMENT

Setting Achievement Levels on the 1994 National Assessment of Educational Progress in Geography and in U.S. History and the 1996 National Assessment of Educational Progress in Science

Final Version

American College Testing
April 1994

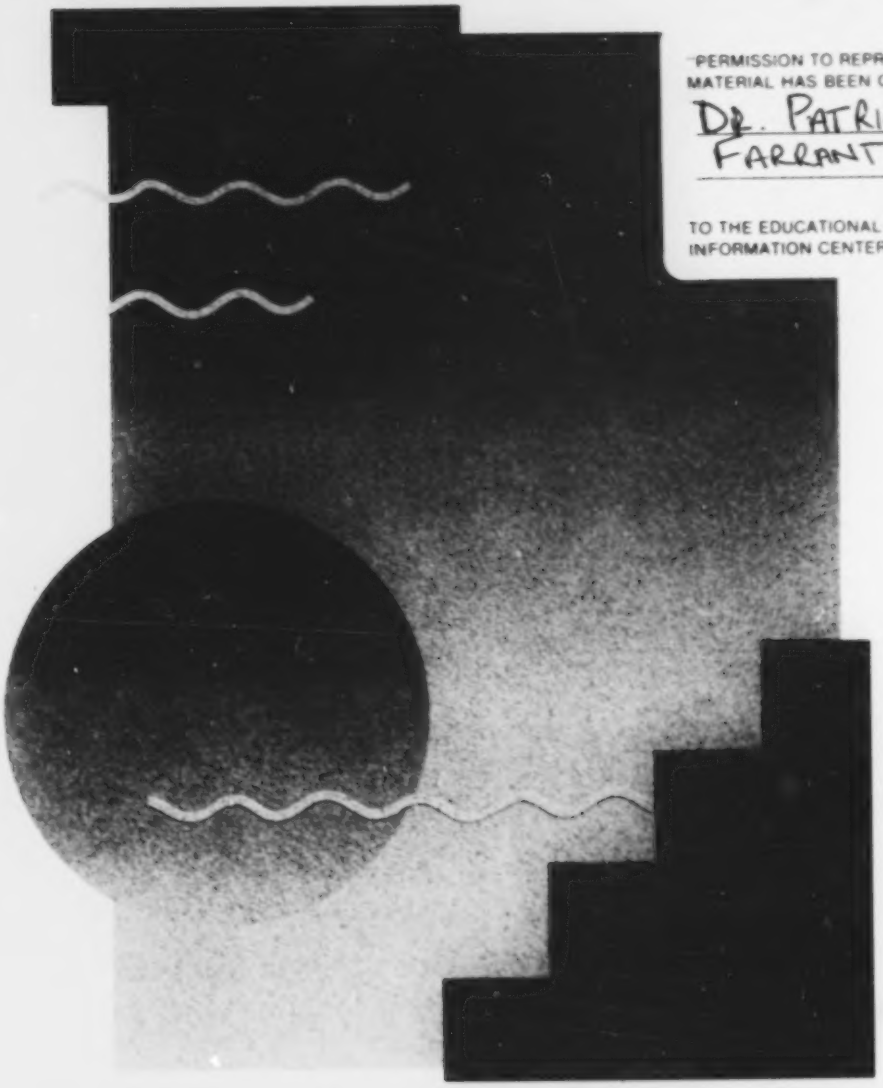
U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Dr. Patricia A. Farrant
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"



So 024 930

BEST COPY AVAILABLE

DESIGN DOCUMENT

Setting Achievement Levels on the 1994 National Assessment of Educational Progress in Geography and in U.S. History and the 1996 National Assessment of Educational Progress in Science

Final Version

American College Testing
April 1994

The work for this report was conducted by The American College Testing Program under contract ZA9003001 with the National Assessment Governing Board.

Copyright © 1994 by The American College Testing Program. All rights reserved.

NATIONAL ASSESSMENT GOVERNING BOARD

MARK D. MUSICK, Chair
President
Southern Regional Education Board
Atlanta, Georgia

PARRIS C. BATTLE
Education Specialist
Office of Grants Administration
Miami Springs, Florida

HONORABLE EVAN BAYH
Governor of Indiana
Indianapolis, Indiana

MARY R. BLANTON
Attorney
Blanton and Blanton
Salisbury, North Carolina

LINDA R. BRYANT
Dean of Students
Florence Reizenstein Middle School
Pittsburgh, Pennsylvania

HONORABLE NAOMI K. COHEN
Former Representative
State of Connecticut
Hartford, Connecticut

CHARLOTTE A. CRABTREE
Professor of Education
University of California
Los Angeles, California

CHESTER E. FINN, JR.
Founding Partner & Sr. Scholar
The Edison Project
Washington, DC

MICHAEL J. GUERRA
Executive Director
National Catholic Education Association
Secondary School Department
Washington, DC

WILLIAM (JERRY) HUME
Chairman
Basic American, Inc.
San Francisco, California

CHRISTINE JOHNSON
Director of Urban Initiatives
Education Commission of the States
Denver, Colorado

JOHN S. LINDLEY
Director, Admin, Training & Development
Clark County School District
Las Vegas, Nevada

HONORABLE WILLIAM T. RANDALL, Vice Chair
Commissioner of Education
State Department of Education
Denver, Colorado

JAN B. LOVELESS
Educational Consultant
Jan B. Loveless & Associates
Midland, Michigan

MARILYN McCONACHIE
Local School Board Member
Glenview, Illinois

HONORABLE STEPHEN E. MERRILL
Governor of New Hampshire
Concord, New Hampshire

JASON MILLMAN
Prof. of Educational Research Methodology
Cornell University
Ithaca, New York

HONORABLE RICHARD P. MILLS
Commissioner of Education
State Department of Education
Montpelier, Vermont

MITSUGI NAKASHIMA
Hawaii State Board of Education
Honolulu, Hawaii

MICHAEL T. NETTLES
Professor of Education & Public Policy
University of Michigan
Ann Arbor, Michigan

HONORABLE EDGAR D. ROSS
Senator
Christiansted, St. Croix
U.S. Virgin Islands

MARILYN A. WHIRRY
12th Grade English Teacher
Mira Costa High School
Manhattan Beach, California

SHARON P. ROBINSON (ex-officio)
Assistant Secretary
Office of Educational Research
and Improvement
U.S. Department of Education
Washington, DC

ACHIEVEMENT LEVELS COMMITTEE

Chair - William Randall
Member - Mary Blanton
Member - Chester E. Finn Jr.
Member - Stephen Merrill
Member - Jason Millman
Member - Michael Nettles

Table of Contents

Introduction 1

Key Points in ACT's Design 4

Section 1—Identification and Selection of Panelists 7
- Stakeholder Input 7
- The Sampling Plan 8
  - The Nominators 9
    - For teacher panelists 9
    - For nonteacher educator panelists 10
    - For panelists from the general public 10
  - The Panelists 11
    - Teachers 11
    - Nonteacher Educators 11
    - General Public 11
  - Sample Size 12
    - For teacher panelists 12
    - For nonteacher educator panelists 12
    - For panelists from the general public 12
- Panelist Selection 13
- External Involvement 14

Section 2—Preparation of Briefing Materials 19

Section 3—Achievement Levels-Setting Procedures 21
- Pilot Study 21
  - Special Features of the Pilot Study 22
  - Research on Informing Panelists More Completely 24
  - Research on Item Rating Methodology for Polytomous Items 31
  - Summary 35
- Implementation 35
  - Orientation to ALS Meetings 36
  - Expanding and Refining Achievement Levels Descriptions 38
  - Training Panelists as Raters 42
  - The Item Rating Process 43
  - Informing the Ratings via Feedback and Other Information 45
  - Recommendations: Cutpoints, Descriptions and Exemplar Items 47
  - Evaluation 48
  - Final Wrap-Up 48

Section 4—Statistical Analyses 59
- Analyses for the Pilot Studies 59
  - Analyses to Support the Rating Process 59
  - Analyses to Determine the Effects of the Experimental Procedures 61
  - Analyses to Determine the Best Procedure for Rating Graded Response Items 62

Table of Contents
(continued)

Potential Limitations of Statistical Analyses for the Pilot Studies 62
Analyses for the Actual ALS Process 63

Section 5—Public Comment Forums 66

References 67

Appendix A Organizations to be Contacted
Appendix B Nominator Materials

List of Tables and Figures

Table 1	Expected Participation of Nominators and Panelists	16
Table 2	1994 ALS Procedures for Pilot Studies	50
Figure 1	Job Titles for Primary and Secondary School Nonteacher Educator Sample	17
Figure 2	Summary of the Nomination Process	18
Figure 3	Draft Agenda for Research Topics and Procedures to be Tested in the Pilot	53
Figure 4	Intrarater Consistency Feedback	64
Figure 5	Interrater Consistency Feedback	65

Introduction

> *Setting appropriate achievement levels on the National Assessment of Educational Progress will help define some of the important outcomes of education, stating clearly what students should know and be able to do at key grades in school. This will make the Assessment far more useful to parents and policymakers as a measure of performance in American schools and perhaps as an inducement to higher achievement. The achievement levels will be used for reporting NAEP results in a way which greatly increases their value to the American Public.*
>
> NAGB 1990

Achievement levels are an important and increasingly integral part of the National Assessment of Educational Progress (NAEP). Achievement levels directly address the function of NAEP to communicate information about student performance in selected learning areas to a variety of constituencies in order to improve education in the United States and to meet goals that signal educational parity, at a minimum, in the international arena. In particular, achievement levels give a readily understood means of describing what students should know and be able to do.

In 1990, the National Assessment Governing Board (NAGB) unanimously adopted three achievement levels to serve as the primary means of reporting results for the NAEP. These three levels are as follows:

> ***Proficient***: Proficient is the central level. It represents solid academic performance at each grade level and competency over challenging subject matter.
>
> ***Advanced***: Advanced is the highest level. It signifies superior performance beyond proficient grade-level mastery.
>
> ***Basic***: Basic is the level below proficient. It denotes partial mastery of knowledge and skills fundamental for proficient work at each grade.

The plan described here is an extension of earlier work by NAGB to set achievement levels on the 1990 NAEP mathematics assessment, and by NAGB and American College Testing (ACT) to set achievement levels on the 1992 NAEP assessments of mathematics, reading, and writing. This project will result in refinements in the process of setting achievement levels as well as in recommendations for achievement levels on the 1994 NAEP in geography and in U.S. history and the 1996 NAEP in science. The experiences of ACT in carrying out the responsibilities of the contract for setting achievement levels on the 1992 NAEP assessment in mathematics, reading and writing inform the design and enhance the innovations to the process.

This design document describes a research methodology for setting achievement levels (performance standards) on the NAEP. The questions of what constitutes U.S. history (or geography, or any other subject assessed in NAEP), how it shall be

assessed, the specific form in which information will be collected, item development, test administration, and so forth are ***all*** questions of importance, but these are "*givens*" to this project. Inputs to this process are:

- Frameworks for the Geography NAEP and the U.S. History NAEP. These frameworks were developed by the Council of Chief State School Officers, under contract to the National Assessment Governing Board, through a consensus process involving panels of experts and public comment forums held throughout the U.S. The geography framework development lasted for eight months, beginning in June 1991. The U.S. history framework development was begun in August 1991 and completed in July 1992.

- Item development and field testing for each subject area was carried out by the Educational Testing Service (ETS) under contract to the National Center for Educational Statistics (NCES). Members of NAGB and the Framework Consensus Panels participated in the selection of items to be included on the assessment in each subject and at each grade.

- Assessments of students throughout the nation selected through a sampling process that samples schools. The assessments in geography and U.S. history were administered during the first three months of 1994. Final assessments (make-ups necessitated by bad weather, for example) were completed by the end of the first week of April. NCES has authority for administration of the NAEP, and ETS develops and administers the assessment under contract to NCES.

- Assessments are scored by National Computer Systems (NCS) under contract to ETS. The scores are reported to ETS where they are converted to the NAEP scale using a three-parameter Item Response Theory Model. The item parameters are provided to ACT for use in setting achievement levels in each of the assessment subject areas.

- Policy definitions of the achievement levels (given above) are set by NAGB. Preliminary achievement level definitions are developed by framework panelists for each grade level within each subject area, e.g. fourth grade Basic geography, fourth grade Proficient geography, and fourth grade Advanced geography.

This process is conducted to set achievement levels on each subject area assessed by NAEP in 1994. This process is designed to produce three products: descriptions of the knowledge and skills that students in each of the three grades assessed in NAEP (4th, 8th, and 12th) should have in order to be classified as performing at each level of achievement, the numerical score (or "cutpoint") associated with that level of performance on the particular assessment administered, and items illustrative of the

kinds of knowledge and skills required of students performing at each level of achievement. Outcomes of this process will be:

- Content-based descriptions of each level of achievement for each of the three grades assessed by NAEP;
- Numerical cutpoints (defining the lower bound at each achievement level) that tie these descriptions to performance on the assessments; and
- Items, available from the 1994 pool of items slated for public release, illustrative of the skills and knowledge characterizing student performance at each of the three achievement levels at each of the three grades assessed by NAEP in these subjects.

The NAEP score associated with the cutpoints and performance at or above each tend to be the focus of attention in NAEP reporting. In addition to deriving recommended cutpoints for Basic, Proficient, and Advanced achievement levels, however, important outcomes of this project are the descriptions of the proposed achievement levels and the sample items and responses selected to represent student performance at each achievement level.

For the first time, the frameworks for the 1994 NAEP assessments include preliminary definitions of the three achievement levels for each grade. As a part of the framework, the preliminary definitions were used to guide the development of the assessment items and tasks. Similarly, these preliminary definitions will guide in the development of the achievement levels to be used for reporting the performance of students on the Assessment. Further refinement and operationalization of these preliminary descriptions will facilitate the rating of items during the achievement levels-setting (ALS) process.

Because relatively few blocks of items in each assessment will be released for public review, the maximum feasible number of items will be selected for consideration as items to represent student performance at each achievement level. These achievement levels descriptions and the illustrative items for each will play prominent roles in communicating the achievement of students on the NAEP.

ACT intends to elicit and engage participation by numerous experts, and interested organizations and individuals. The final product will benefit from the input of these many individuals and interests. ACT will provide the impetus for the accumulation of information and expertise to be focused upon and channeled into the development of the achievement levels and the validation of their interpretations of student performance.

The achievement levels will be developed by a group of individuals selected to be representative of both the educational community and the general public. A broadly representative set of panelists will be identified for each of the three content areas.

In all phases of this project, the involvement and participation of stakeholder groups and other interested constituencies will be elicited. Further, the recommended achievement levels—descriptions, numerical values, and illustrative items—will be made available for public review, and an intensive campaign will be launched to engage the public in this review. All comments offered during this public review phase will be shared with NAGB, and these comments will become a part of the information compiled to develop the recommendations for NAGB regarding achievement levels in each of the content areas.

Key Points in ACT's Design

ACT has extensive experience in assisting major national organizations in determining criterion scores or standards for their programs. ACT has knowledge of current advances in achievement levels-setting processes, and the creativity and expertise to modify and expand upon these as appropriate for specific implementations. In addition, ACT has the experience of having designed and implemented the achievement levels-setting process for the 1992 NAEP in mathematics, writing, and reading. We believe that this experience provides us the insights and understanding, merged with the technical and methodological expertise and experience, for designing a process to fully address the many and varied challenges and requirements of the NAEP assessments.

In designing this process, ACT has carefully reviewed the procedures implemented in setting achievement levels for the 1992 NAEP in mathematics, reading, and writing. We have attempted to evaluate objectively the major features of that process and to identify ways to improve upon that process as well as ways to incorporate additional requirements for setting achievement levels on the 1994 NAEP in U.S. history and geography, and the 1996 science NAEP.

Many features of the 1992 process have been retained; many have been improved upon and enhanced. A sampling plan for identifying and recruiting panelists was successfully designed and implemented, and panels of broadly representative, qualified individuals participated in the 1992 ALS process for mathematics, reading, and writing. The approval of a diverse set of interested individuals, organizations, and groups was sought and won for both the sampling plan and the overall research design. The process implemented in 1992 incorporated state-of-the-art methodologies in standard setting, and fully accomplished the goals required of a successful standard setting process. Questions and concerns emerged and were raised regarding psychometric and standard setting issues that had never before been addressed. ACT raised several of these, and we openly and frankly addressed all that were brought to our attention. The design presented in this document incorporates improvements and enhancements generated through the experiences gained during the previous achievement levels-setting efforts for the NAEP.

Key features of the 1994 proposal include the following:

1. A *sampling plan* for recruiting panelists for each achievement levels-setting (ALS) meeting that will result in the involvement of a well qualified, representative panel of judges while introducing efficiencies resulting from the experiences of identifying and recruiting panelists to the 1992 panels;

2. A *research agenda* incorporated into the basic design of the ALS process and validation process that will contribute significantly to the product of the 1994 achievement levels-setting process *and* to the body of knowledge in standard setting, item response theory (IRT), and other technical and methodological areas of educational assessment and measurement;

3. *Pilot studies* for each content area in the 1994 NAEP assessment and for the 1996 science NAEP that will provide the opportunity to test the process and make needed changes and adjustments, as well as provide the opportunity to collect data for carrying out needed research;

4. Extensive *training for panelists*, not only in the methods of evaluating items and rating them to set achievement levels, but also in the consequences of those standards, in the goals and purposes of NAEP and NAGB, and in educational assessment issues and policies;

5. *Ample time* in the agenda for the achievement levels-setting pilot studies and meetings for the key elements of the achievement levels-setting process to be accomplished responsibly and successfully;

6. *Customized computer software* that utilizes the IRT calibrations of the NAEP items and scale to produce feedback, on site, to panelists on the consistency and convergence of ratings, and on the consequences of their ratings;

7. The *same, well-trained staff* for each pilot study and each achievement levels-setting meeting to ensure consistency in implementation;

8. *On-site logistic planning and support services* using full-time, ALS-experienced project staff; and

9. A team of veterans represented on the project staff, the internal advisory team, and the external committee of technical advisors including *highly experienced experts* in standard setting methodology, psychometrics, sampling statistics, educational assessment, collective decision making, and meeting management.

American College Testing's general approach to deriving recommended achievement levels for the NAEP is guided by five overarching principles:

1. there must be broad, thorough, and open participation by all relevant populations in the levels-setting process;

2. highly sensitive and confidential materials, reports, and information must be handled in an appropriate manner;

3. the levels-setting process must be carefully designed, technically sound, rigorously implemented, and appropriately validated;

4. the levels-setting process must be comprehensible to interested parties and easily implemented by process participants; and

5. NAGB must exercise informed direction over all major project activities and be kept fully apprised of all relevant project information.

Section 1—Identification and Selection of Panelists

Because the achievement level recommendations must be derived from the collective judgments of many important audiences, it is essential that the Achievement Levels Panels be carefully selected and broadly representative. In addition to the primary requirement that the panelists be competent to perform the tasks involved in the ALS process, both demographic characteristics and group size are key considerations in the selection of panelists.

NAGB has specified that the composition of panels is to be broadly representative, and that 70% are to be educators and 30% non-educators. Moreover, classroom teachers should comprise 55% of the group. ACT will empanel three groups of 30 members (each for Grades 4, 8, and 12) for each of the three content areas. Thirty panelists per group would allow each group to contain sixteen or seventeen (55%) teachers, four or five (15%) nonteacher educators, and nine (30%) non-educators. Having at least nine achievement levels panel positions reserved for non-educators ensures that important perspectives outside education will be represented.

ACT plans to implement the same basic design implemented in 1992 again for selecting panelists to set achievement levels for 1994 (and future) assessments. This means that all prospective panelists (educators and non-educators) must be familiar with the knowledge and skills required by the content area panel and grade level group to which they are nominated. The criteria recommended to nominators of each type of panelist are detailed below. Requiring comparable, relevant background experiences among panelists bolsters the validity of the process and contributes to greater group cohesiveness. Further, it is ACT's belief that the best way to meet the intent of NAGB's policy calling for achievement levels recommended by a "broadly representative group of judges" (NAGB, 1993, p. 10) is to form a common will from a panel of persons who meet the distributional requirements of NAGB policy and who have practical experience with and knowledge of students at the specified grade levels.

Stakeholder Input

Because many stakeholders are involved in this process—those with great and sincere interest in the outcome of this effort—we believe it is important to have their input in the sampling design and panelist selection aspect of the 1994 ALS process. To that end, ACT has distributed a draft of the ALS Design Document to approximately 200 national organizations and to other groups that have an important role in education or in the different content areas. The draft detailing the panelist selection design, pilot studies, and the procedures planned for implementation in the ALS process was mailed along with an invitation to meet with ACT to discuss concerns regarding the plans for the 1994 ALS process. The first letters were sent on January 26, 1994 to notify the stakeholders that they would soon receive a draft version of the Design Document. That mailing included a form for signing-up to meet in Washington or for

agreeing to send written comments. The excerpted copy of the draft Design Document was mailed to these stakeholders on February 8.

A total of 19 persons signed up for meetings in Washington, and 15 attended. (Please refer to Appendix A for the list of organizations identified as stakeholders for this purpose.) Comments from those attending the meetings were generally quite positive. Some expressed concerns, however, and those concerns and other suggestions for change are included in Appendix A. Written comments have been received from 11 people, and 40 respondents indicated that they would send written comments; 27 indicated regrets that they would be unable to provide input at this time, but requested to be kept informed about the project; and 11 people expressed no interest in the project at this time.

The total response is approximately 50%, which we think is quite positive. One general message from the stakeholders who participated in the meetings was that the document was too dense, too filled with jargon, and too technical to communicate well with them. ACT's Project Director promised to send them "user friendly" document that summarizes the project. That document has been prepared and will soon be ready for distribution to all stakeholders.

The Sampling Plan

School districts will serve as the basic unit of sampling. School districts in U.S. territories will be included in the sampling frame. Three samples without replacement will be drawn from the MDR[1] database of school districts. One sample of 130 districts will be drawn to identify nominators of teachers, one sample of 15 districts will be drawn to identify nominators of nonteacher educators, and one sample of 100 districts will be drawn to identify nominators of the general public. The samples will be drawn to provide roughly equal representation of the four NAEP regions for which the most recent census data (1990) show the following population distribution:

Northeast	20%
Southeast	26%
Central	24%
West	29%

In addition, 15% of the districts will have student enrollment of 50,000 or more students because approximately 15% of the students in the U.S. are enrolled in districts enrolling 50,000 or more students. This dichotomous enrollment variable will be selected over a rural/urban classification. The rural/urban dichotomy obfuscates the important "suburban" community type but a three-way criterion

[1] The name MDR refers to a computer file of school information maintained by Market Data Retrieval, Inc., of Westport, Connecticut.

variable is not feasible from a statistical perspective. The enrollment variable has the added benefit of assuring representation of the Great City Schools districts.

Socio-economic status (SES) is an important indicator of educational policy positions from opportunity-to-learn issues to expenditures. Since the MDR includes a variable indicating districts having at least 25% or more of their population living below the poverty level, this indicator will be used. Analysis of the district data indicates that approximately 15% of the districts had 25% or more of their population classified as being below the poverty level, and 15% will be taken as the target for "low SES" districts to be included in the sample.

Since private school representation is also desired, two samples without replacement will be drawn from a database of private schools. One sample of 33 schools will be drawn to identify nominators of teachers, and the other (with only 5 schools) to identify nominators of nonteacher educators. The samples will be drawn to provide roughly equal representation of the four NAEP regions based on the most recent data on enrollment in private elementary and secondary schools (NCES, 1993):

Northeast	31%
Southeast	19%
Central	28%
West	22%

Finally, a systematic random sample of higher education institutions will be drawn from the *1994 Higher Education Directory* (published by Higher Education Publications, Inc.) to identify nominators of nonteacher educator panelists.

▪ ***The Nominators***

In order to involve a variety of interests in the process of nominating and recruiting panelists, a large and diverse set of nominators will be contacted. The nominators include principals, superintendents, school board presidents, leaders of district teachers' associations, nonteacher educators at the secondary and postsecondary levels of education and at the state level of jurisdiction, local chief elected officials, and local civic leaders with education interests. (See Figure 1 column 3 for different types of nominators.)

For teacher panelists. For the districts drawn for the teacher nominators the school district superintendent and the head of the bargaining and/or largest teachers' organization will be asked to nominate teachers from their districts. The curriculum supervisor for each state included in the teacher district sample will also be asked to nominate teachers. The principals or superintendents of private schools will be asked to nominate teachers from those schools. Each district superintendent, association president, and private school principal or superintendent will be asked to nominate up to four individuals from each of the three grade levels from among the teachers in his or her district or school who meet the criteria for nomination. The state

curriculum supervisors will be asked to nominate up to four teachers from each of the three grade levels from any district in his/her state.

All nominators will be asked to keep in mind the need for appropriate distributions of gender and ethnicity when making their selections, and to report the sex and ethnicity of each of their nominees. Finally, these nominators will be asked to permit (or secure permission for) any nominees who are selected as panelists to attend the ALS meetings. (See the sample letter to teacher nominators in Appendix B.)

For nonteacher educator panelists. The nominators of nonteacher educators will themselves be nonteacher educators. Postsecondary nominators can be teachers, too, but they qualify as nonteachers because "teacher" refers to K-12 classroom teachers. The nominators may nominate themselves or any of their colleagues who meet the specified requirements.

Four sources will be used to obtain nominations for nonteacher educators. The nominators for the primary and secondary school district level and private school nonteacher educators will be selected from the "MDR Personnel File" according to job title. The job titles included in the sample are listed in Figure 1. The nominators for the state-level nonteacher educators will be a state-level education officer (either the commissioner/chief, assessment director, or curriculum director) from each state included in the nonteacher district sample. Either the Dean of Liberal Arts, Dean of the College (for Liberal Arts Colleges), Dean of Instruction (at community colleges), or Education at each two-year and four-year postsecondary institution sampled will be asked to serve as a nominator in this category, as well.

The nominators will be requested to nominate from one to four individuals for each of the three grade levels in a particular subject area. They will also be requested to keep in mind the distribution of sex and ethnicity among their colleagues and to provide information on these variables for the people they nominate. (See the sample letter to nominators of nonteacher educators in Appendix B.)

For panelists from the general public. Nominations for the general public panelists will be obtained from these different groups: 1) the Chair of the Education Committee of local Chambers of Commerce; 2) mayors (or equivalent level of elected official) of local municipalities; and 3) chairs of district school boards. Nominations from these three groups may include themselves, if they meet the criteria, but the general public category is not restricted to members of the Chamber of Commerce, mayors nor school board chairs. Representation of the general public does, however, exclude the educational community. Nominators will be specifically instructed not to nominate former teachers and educators, in order to ensure that this sample does represent the non-educational community. (See the sample letter in Appendix B to this group of nominators.) Each nominator will also be requested to provide information on the sex and race/ethnicity of the individuals nominated.

ACT will use the methodology previously described to draw a sample of districts for identifying persons to nominate panelists from the general public. Names and addresses will be obtained by telephoning the district superintendent's office for information or the office of elected public officials (mayors), Chamber of Commerce, and School Board, and so forth, when necessary.

Once the nominators are identified, they will be contacted in writing and asked to nominate up to four individuals at each grade level in the relevant target group of panelists, and to report the sex, the race/ethnicity, and other information for each nominee.

▪ ***The Panelists***

ACT is committed to identifying and selecting panelists who are informed and knowledgeable about the content area and are reasonably "current" in their familiarity with what school students are expected to know and do at the relevant grade level (4th, 8th, or 12th). To this end, nominators for each target panelist group will be asked to nominate individuals who meet the necessary qualifications. We have been as precise as possible about defining the target groups from which panelists will be selected. This precision is required to meet sampling assumptions.

Teachers. Panelists nominated to the pool must meet all of the following qualifications:

a. At least five years of overall teaching experience.
b. At least two years of teaching experience in the subject matter and with students in the indicated grade (4, 8, or 12).
c. Judged to be "outstanding" in their professional performance by a supervisor or someone in the position to make that judgment.

Nonteacher Educators. Panelists will be nominated by and from three groups of nominators:

a. Nonteacher educational staff at primary and secondary educational institutions.
b. Selected positions in state departments of education.
c. Professors or administrators at postsecondary institutions.

Panelists nominated from any nonteacher educator group must have familiarity and professional experience with the subject matter of the test at the indicated grade level, and must be judged "outstanding" in their professional performance by the nominator. The nominator will also be asked to indicate the reason for which the person is considered to be outstanding.

General Public. Persons nominated from the general public to be panelists must

a. Have familiarity with the content area at the indicated grade.
b. Not have been employed by an educational institution in the past. For example, a parent of a fourth-grade student and an employer of recent high school graduates might qualify as members of the general public

target population; but, a former teacher, principal, or district superintendent would not qualify.

In addition to the information from each nominator indicating the reason for which each nominee is considered to be outstanding, prospective panelists will be interviewed, *via* telephone, to verify their credentials.

■ ***Sample Size***

The sizes of the samples that will be drawn are based mainly on the 1992 experience. Information from the 1992 experience used to compute the sample sizes includes the number of nominators identified for each district sampled, the response rate of nominators, and the approximate number of individuals nominated by each nominator. The acceptance rate for invited nominees who agreed to participate in the ALS process was also considered. Information from the 1992 reading panels recruitment process indicates that:

For teacher panelists:

1. Approximately 1.70 nominators were identified for each district in the sample.

2. Approximately 36% of the nominators who were contacted responded with at least one nomination.

3. Approximately 3.14 individuals were nominated by each nominator.

4. Approximately 78% of nominees invited to serve as panelists agreed to participate in the ALS process.

For nonteacher educator panelists:

1. Approximately 1.58 nominators were identified for each district in the sample.

2. Approximately 31% of the nominators who were contacted responded with at least one nomination.

3. Approximately 4.31 individuals were nominated by each nominator.

4. Approximately 91% of nominees invited to serve as panelists agreed to participate in the ALS process.

For panelists from the general public:

1. Approximately 2.22 nominators were identified for each district in the sample.

2. Approximately 18% of the nominators who were contacted responded with at least one nomination.

3. Approximately 4.00 individuals were nominated by each nominator.

4. Approximately 65% of nominees invited to serve as panelists agreed to participate in the ALS process.

The expected number of nominators and panelists are summarized in Table 1. Three samples of school districts will be drawn for each subject area: 130 districts for nominators of teachers, 100 for general public representatives, and 15 districts for nominators of nonteacher educators. Two samples of private schools will be drawn for each subject area. The samples will include 33 private schools for nominators of teachers and five private schools for nominators of nonteacher educators. The sample of districts will be stratified by region, community type, and student enrollment size. The sample of private schools will be stratified by region only. In addition, a sample of 15 universities and colleges will be drawn to identify nominators for the nonteacher educators. (See Figure 1 for a chart summarizing the nomination process.)

Based on the assumptions, the above samples will permit us to invite one in every four nominees, and have approximately 50 teacher panelists, 14 nonteacher educator panelists, and 26 panelists representing the general public, for each content area. Moreover, it is expected that about 10 panelists will be selected from private school nominations.

Panelist Selection

While the method of selecting samples of school districts, private schools, and postsecondary institutions represents "probability sampling" in which each member of a well-defined target population has a known positive probability of being selected, the selection of panelists is *not* probability sampling *per se*. Probability sampling at the panelist level is not possible because of the unknown and subjective judgments of nominators. The *process* can be replicated, and there is no particular reason, given previous experiences, to believe that results of this sampling process would greatly differ. By using aspects of sampling methodology, however, we will be able to select broadly representative panels through which diverse points of view can be expressed.

Each individual in the pool of nominees for each content area will be categorized using eight variables:

Grade Level	Panelist Type
Grade 4	Teacher
Grade 8	Non-Teacher Educator
Grade 12	General Public

Race/Ethnicity
- White
- Black
- Asian
- Native American
- Hispanic

Gender
- Male
- Female

Region of Nominator
- Northeast
- Southeast
- Central
- West

Community Type of Nominator
- Low SES
- Not Low SES

District Size of Nominator
- <50,000
- ≥50,000

School Affiliation
- Public
- Private

Individuals from the pool of nominees will be selected and invited to participate. Thirty individuals will be empaneled for each of the three grade levels (4th, 8th, and 12th). In order to meet the NAGB policy, for each grade level 16 or 17 (55%) of the panelists will be teachers, 4 or 5 nonteacher educators (15%), and 9 (30%) non-educators. Considering the small panel size, it will not be possible to ensure that each panel is representative with respect to each combination of characteristics (*i.e.*, Hispanic females in small districts in the central U.S.). Moreover, it will not be possible to ensure proportional representation of categories. However, to the extent possible, panelists will be selected from the pool of nominees so as to maximize the balance of gender and race/ethnicity, as the primary considerations, and geographical region, school affiliation (public/private), type of community (socio-economic status), and district enrollment size, as secondary considerations (each of equal weight). While this does not ensure proportional representation among different criteria, it does ensure diversity among the members selected for the panels.

External Involvement

In order to allow interested groups and organizations to have input in the nomination of panelists, the lists of nominators will be distributed to key professional organizations and groups. The lists will be distributed as soon as the nominators are identified. The tentative schedule for distribution is as follows:

Pilot Studies
U.S. History: April 12, 1994
Geography: April 12, 1994
Science: February 3, 1996

ALS Meetings
U.S. History: August 24, 1994
Geography: August 24, 1994
Science: February 17, 1996

The organizations may **not** submit nominations, but they may wish to influence the selections made by nominators.[2] Examples of appropriate uses of these lists are to contact nominators for the purpose of urging them to make nominations, to urge them to make nominations on the basis of any criteria the group feels is particularly important in their interests *and* consistent with the criteria outlined above, and to *suggest* persons to the nominator that he/she should consider.

ACT believes that the level of participation described above allows interested organizations and groups an opportunity to be a part of the process of identifying panelists, but it does not provide a level of influence that would significantly "stack" the panels. By promoting and encouraging broader participation in the nomination process, ACT believes that the pool of nominees will be sufficiently large, qualified, and motivated to assure an outstanding set of panelists. Moreover, ACT is confident that these initiatives will enable NAGB to reach all major segments of the education community, and that these initiatives will facilitate the development of achievement levels that have taken into consideration the views of the broadest possible spectrum of society.

[2] Geisinger (1991) suggests that panelists with personal stake in the outcome of standard setting should be eliminated, and that ratings from judges who do not give independent ratings should be eliminated.

Table 1

Expected Participation of Nominators and Panelists

Type of Panelists	Sample Sizes	Nominators Identified	Nominator Response	Nominees	Invited Nominees	Panelists
Teachers	130 districts 33 private schools	228[1]	82[1]	258[1]	64	50
Nonteacher Educators	15 districts 5 private schools 15 colleges/universities	47	15	16	16	14
General Public	100 districts	220	40	160	40	26
All	245 districts 38 private schools 15 colleges/universities	495	137	434	120	90

[1] State Social Studies Curriculum Directors were not included as nominators of teacher panelists in the 1992 ALS process; therefore, the response rate for this group is not known. Nor is it known how many different states will have been identified in the sample of districts. This count, therefore, does not include the State Social Studies Curriculum Directors.

Figure 1

Job Titles for Primary and Secondary School Nonteacher Educator Sample

K-12 Curricular/Instructional Supervisor
Elementary Curricular/Instructional Supervisor
Secondary Curricular/Instructional Supervisor
K-12 Guidance Counselor/Supervisor
Elementary Guidance Counselor/Supervisor
Secondary Guidance Counselor/Supervisor
K-12 Social Studies Supervisor
Elementary Social Studies Supervisor
Secondary Social Studies Supervisor
K-12 Science Supervisor
Elementary Science Supervisor
Secondary Science Supervisor
Elementary Principal
Secondary Principal
Assistant Principal
Admissions Director
Assessment Coordinator

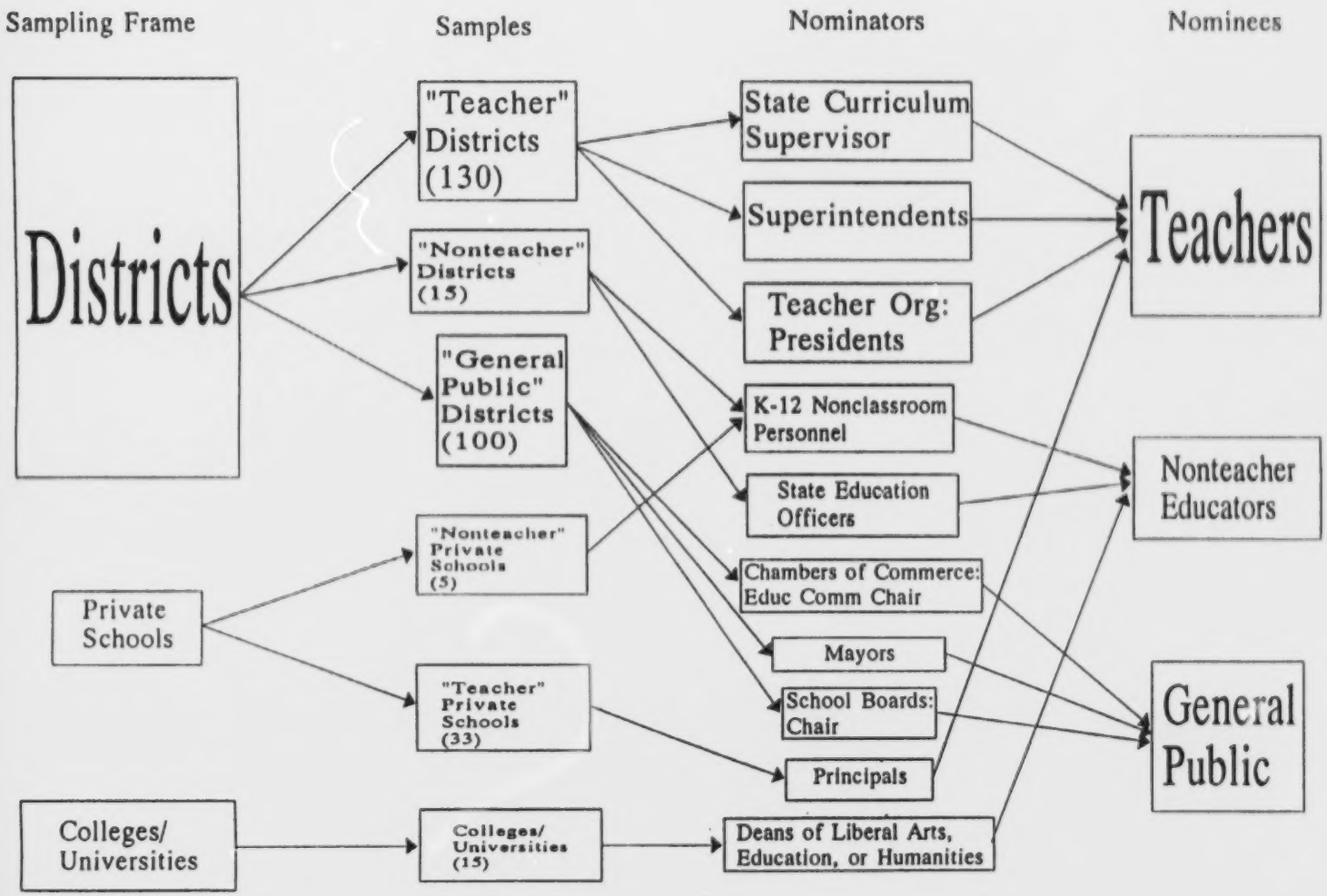


Figure 2
Summary of the Nomination Process

Section 2—Preparation of Briefing Materials

One of the most critical elements of a successful achievement levels-setting procedure is that participants be thoroughly familiar with the methodology to be employed and have a sufficient understanding of key background materials. Published research provides little guidance on this topic (Mills, Melican, & Ahluwalia, 1991). For example, the profession's Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985, p. 43) note only that the qualifications of participants should be documented. Recently, researchers have called attention to the fact that the various backgrounds and experiences of participants are a source of concern (see Plake, Melican, & Mills, 1991; Reid, 1991), but *a priori* interventions to ameliorate the concern are rarely offered. Background characteristics and criteria for selection of panelists were discussed in the previous section.

ACT's general approach to preparing the panelists for the task of setting achievement levels for the NAEP will be to provide early, pre-meeting intervention. Through many years of refining successful standard-setting procedures, ACT has developed an approach that pre-empts the often-encountered difficulties of panelists arriving at a meeting unprepared for the task. ACT will provide all panelists with a set of materials that briefly, yet fully and comprehensibly, explores important background topics and provides introductory information regarding the achievement levels-setting task. The latter information is especially critical because the judgmental levels-setting methodology is often new and challenging to meeting participants. For other projects, ACT has succeeded in developing interesting materials that employ intuitive strategies for communicating with persons new to judgmental levels-setting procedures. It is our goal to provide just this sort of preparatory materials for panelists involved in the NAEP achievement levels-setting process. To that end, we will also consult with organizations such as the National Council for Geographic Education, the National Council for History Education, the National Council for Social Studies, and the National Science Teachers Association for suggestions and recommendations of materials to distribute to panelists.

The successful implementation of the levels-setting methodology begins at the very outset of the project. As just stated, ACT will provide informative materials to panelists at least two weeks in advance of the meeting date to allow them to become thoroughly familiar with the materials. Materials provided to panelists will include the following:

1. cover letter, prepared in consultation with the NAGB's Achievement Levels Committee;

2. statement of the NAEP mission, principles, and objectives;

3. security agreement;

4. NAEP Overview, "The Test for Our Society";

5. meeting agenda, hotel accommodations confirmation, etc.;

6. brief introduction to the judgmental levels-setting methodologies to be used in the project;

7. the appropriate NAEP framework document;

8. brief description/discussion paper on topics such as the circumstances for administration of the NAEP, item formats and differential student performance on the different formats, and student test-taking behaviors;

9. copy of NAGB's policy framework, "Setting Appropriate Achievement Levels for the National Assessment of Educational Progress";

10. copy of the "user friendly" Design Document; and

11. referral sources of information for participants who have questions about the achievement levels-setting materials.

Combined, such a set of common reference materials and intuitive, introductory information related to the levels-setting process should form a sound basis for achieving a successful achievement levels-setting procedure.

ACT will obtain these briefing materials and other data/information for the ALS meetings from ETS (the Operations Contractors for NAEP) and their scoring subcontractor NCS, the National Center for Education Statistics (NCES) and NAGB. We will develop any additional materials needed to inform and train the panelists.

Section 3—Achievement Levels-Setting Procedures

The proposed design of the achievement levels-setting process consists of several phases: 1) a pilot study to collect research data in support of the future ALS processes and to try out the basic design; 2) implementation of the procedures; and 3) evaluation of the process and results *via* validation studies, public comment forums, consultative sessions with stakeholder groups and organizations, meetings of TAT and TACSS, and meetings with the NAGB staff, Achievement Levels Committee, and Board. The first two of these phases are described in this section.

Pilot Study

It is vitally important for those planning an achievement levels-setting procedure to know *a priori* the ways various elements interact with each other in the achievement levels-setting process. For example, Achievement Levels Panelists interact with the pre-meeting materials, the meeting materials (i.e., the test questions, rating forms, rater feedback, and so forth); each other, and with the project staff. All of these elements combine to promote or degrade what has been called "intrajudge consistency" and "interjudge consensus" (Friedman & Ho, 1990).

Previous research has conceptualized the effects of two major kinds of interaction: 1) people interacting with text (Smith & Smith, 1988), and 2) people interacting with each other (Curry, 1987; Fitzpatrick, 1990). In order to assess the effects of textual and social interaction and adjust study procedures accordingly, ACT proposes that a pilot study of the achievement levels-setting process for both U.S. history and geography be included as the first phase of the project. The pilot study for science will be conducted in April 1996, assuming that science is assessed in 1996.

The pilot study for geography will be conducted July 14-18, 1994, and the pilot study for U.S. history will be August 11-15, 1994. Each pilot study will use the same design, location, procedures, and materials proposed for the actual levels-setting meeting. Only field trial data will be available to use for the pilot studies, however. This will pose some problems, especially with respect to item parameters for items that were changed for the 1994 NAEP. The data have been evaluated, and ACT is aware of several problems. The timeframe for reporting, however, necessitates that the pilot studies be conducted before the operations contractor can have the operational data available for our use.

The pilot studies will be used to collect data for some issues that were identified during the 1992 process, as well as for some innovations to be evaluated for implementation in the 1994 process. These pilot studies will be conducted under the general principle that collecting information about the procedures in the process to be implemented is more important than collecting precise numerical data on achievement level cutpoints. Thus, when a choice must be made between implementing a procedure and collecting "clean" ratings, the choice will be for implementing the procedure. The goal is, of course, to maximize the amount of

information obtained both for the ALS procedural outcomes and the ALS numerical outcomes.

Pilot study panelists will be selected for each of the three grade groups. They will be selected using the same nomination and selection procedures described in Section 2 and according to the same criteria as described for selection of the actual ALS panelists (i.e., balanced by gender, ethnicity, vocation, etc.). The only difference in selection of panelists for the pilot studies and the ALS meetings is that only 20 persons per grade group will be impaneled for the pilot study, as opposed to 30 each for the ALS processes.

The number and complexity of items increases with increasing grade levels. In U.S. history, there are 94 items in the pool for Grade 4, 148 for Grade 8, and 156 for Grade 12. In geography, there are 90, 125, and 123 for the three grades, respectively. While the distribution of item formats across the three grade levels is relatively equal, the absolute number of extended response and performance tasks is greater at the higher levels. The significance of this for timing became obvious during the implementation of the 1992 process. Only one pilot study was conducted in 1992, and that was for Grades 4 and 8 only. The pilot studies currently planned allow us to determine how much time is needed by panelists in the different grade levels during each step in the process for each content area.

General public panelists will be asked during the pilot studies whether they should have additional training. If they indicate that additional training is needed, ACT will bring general public panelists to the meeting site early, in order to begin training.

Further, the grade levels at which these subjects are offered in schools do not correspond perfectly with the grade levels assessed by NAEP. The pilot studies provide an opportunity to determine whether and how this incongruence may effect the ALS process—overall, by grade level, by subject area, or by particular step(s) within the process. At present, the plan is to request that nominations of teacher panelists be restricted to persons who teach the subject (e.g., geography) and who teach students at the grade level (4th, 8th, or 12th) for which the teacher is being nominated to serve as a panelist.

▪ ***Special Features of the Pilot Study***

Pilot study panelists will receive the same background materials that will be sent to panelists in the actual achievement levels-setting meeting in each content area. Every aspect of the process to be implemented for setting achievement levels will be tried out in the pilot study. The pilot studies will, however, include some additional steps and procedures that are to be tried out and evaluated to determine whether and how they might be incorporated into the ALS process. The pilot studies are ***not*** designed nor intended to address questions of replicability with respect to the procedures to be implemented in the actual ALS process

Panelists in each grade group will be divided into two groups, and each group will rate approximately one half of the items in the item pool for that grade. Items are grouped by blocks, and there are approximately 15 items in each block timed for 25 minutes. Items will be divided into rating pools by blocks so that both halves of the rating pool are as equal as possible with respect to item difficulty (indicated by student performance data), item formats (multiple choice, short answer, and extended constructed response or performance tasks), content specifications (e.g., items assessing knowledge and skills related to space and place, environment and society, or spatial dynamics and connections), or items and tasks employing supplementary tools and devices (atlases, seed packets, compasses, and so forth). Previous experiences with developing matched halves of the item pools for each grade were quite successful.

Dividing panelists into two equal groups is not so easy, however. The requirements for representation were detailed in Section 1. It is not possible to have exactly equal numbers of teachers, nonteacher educators and general panelists in the two groups because an odd number of panelists will represent some or all of these categories. The same will be true for persons representing the various demographic attributes to be represented on the panels. In particular, many more females than males teach at lower grade levels, in general; and the ratio of males to females in teaching, education-related professions, and the general public varies greatly according to the discipline/curricular area assessed. One expects to find males more equally represented in geography, history, or science than in reading or writing. Nonetheless, sex, race/ethnicity, region, panelist type, and so forth can only be approximately equally distributed between the two item rating groups of panelists. These rating groups will be used for purposes of experimental controls in the pilot studies to be conducted for the 1994 ALS process.

ACT believes it is important to get information about how panelists conceptualize performance for achievement levels. This question is related to the issue of whether item-by-item ratings, *versus* a more holistic approach, accurately reflect panelists' evaluations of student achievement. Panelists will be asked about their conceptualization of student performance with respect to the magnitude of evidence required to characterize a student at the "Basic," "Proficient," or "Advanced" level.

This question arose with respect to ratings for the 1992 Writing NAEP. All tasks required extended constructed responses. Two methods of mapping ratings from the paper selection process were examined during the evaluation of those data. One method used plausible values (generated by ETS for scaling purposes), and the other used information from the test characteristic curve (TCC). The TCC methodology effectively implies that "Advanced" means advanced in *all* types of exercises or tasks included on the assessment, whereas the plausible values method effectively implies that "Advanced" means that the student can demonstrate this level of performance in *perhaps* only one type of exercise or task. Whether the student can consistently perform at the Advanced level is not taken into account with the plausible values methodology. The key question that could not be answered in 1992 was which

conceptualization was guiding the panelists in their choice of papers to represent student performance at each achievement level. Ascertaining how panelists conceptualize student achievement during the 1994 pilot studies will help shed light on this issue.

▪ ***Research on Informing Panelists More Completely***

ACT proposes to provide panelists with information targeted at anchoring their ratings in *reality*. Both interjudge and intrajudge consistency were quite high in the procedures implemented for the 1992 NAEP assessments in mathematics, reading, and writing. It was not clear, however, whether ratings resulted in distributions along the NAEP scale that would be acceptable to panelists, had they been made aware of the consequences of their ratings with respect to the empirical distributions of student scores on the NAEP assessments. For example, it was not clear that panelists—even those who had high levels of internal consistency and who had rated items consistently with other panelists—would have been in agreement with setting an achievement level that resulted in as few as 1% of students performing at or above the advanced level or as many as 20% at that level.

To get more information on this sort of question, ACT will collect information and provide feedback to panelists that will help determine whether panelists find the consequences of their item ratings acceptable. This decision will be guided, in part, by the panelists' responses to direct questions regarding the utility of the information during the ratings process.

a. Performance Distributions

Panelists will be asked to provide estimates of student performance at or above the achievement levels, as defined previously in the process. They will be asked to estimate the percentage of students that would score at or above each achievement level on the NAEP. Panelists will be instructed on the meaning of "at or above each achievement level. For example, "at or above Basic" includes students scoring at the Proficient and Advanced level too. "At or above Proficient" includes students scoring at the Advanced level as well as at the Proficient level. A chart or graph will be prepared to demonstrate this point in the training/instructional presentation of this task. They will have been instructed on matters relating to NAEP testing conditions, NAEP reporting practices, and so forth. They will understand that student performance on the NAEP does not necessarily conform to student performance on "high stakes" assessments.

The "actual" percentage will be computed, based on the aggregate of the item-by-item ratings for each achievement level. The difference in percentages estimated by panelists and the empirical (estimated) NAEP score distribution percentages, computed from panelists' ratings, will be presented to panelists. The ratings data will be used after each round to compute the cutpoint and estimate the percentage of students that would score at or above the cutpoint for each achievement level. Before Round 3 ratings, panelists in Item Rating Group A for each grade will be given consequences feedback information. They will be given data on the percentages of

students scoring at or above each achievement level for their grade, based on the achievement level cutpoints that would result from the Round 2 ratings. The individual estimates of the percentage of students who would score at or above each achievement level will be averaged to report to the panelists as well. They can compare the "actual" percentages, given current cutpoints, to estimates percentages.

These aggregate data will inform panelists about the consequences of their ratings and indicate whether adjustments are advisable. The purpose of providing panelists with these data will be fully explained to *all* panelists before the first round of ratings. This message will be reiterated to the Group A Rating Group when the data are provided to them. Panelists who are participating in the experiment using consequences feedback data will be instructed in the importance of *not* sharing this information with the other "control" group panelists who will not have this information.

For purposes of sharing this information with panelists prior to Round 3 ratings, the "consequences treatment" group will be convened separately from the other group. Then, in the grade level group, they will be given data on the achievement level cutpoint computed for each group alone, and they will be given interjudge consistency data plotting only ratings for members of the treatment group and plotting only ratings for the control group.

Panelists' estimates of the percentages of students at or above each cutpoint can be used to recover the "percentage correct" estimate for each panelist that would be associated with their estimate of the percentage "at or above" that cutpoint. Distributions of panelists' estimates on the theta (or a theta-like) metric can be represented on graphs just like those used to present interjudge consistency feedback. This will allow each panelist to see their estimates, relative to those of other panelists in their Item Rating Group. Further, panelists can see how their own estimates relate to their own item ratings. Panelists will be asked to comment on the differences between the average estimated by the group and the average computed from item ratings. (Questions on the evaluation instruments administered each day about each round of ratings will be developed to collect this information.)

The rationale for collecting this information is to codify the "what is" versus "what should be" distinction. Since panelists must make any adjustments to the overall grade-level standard via their individual item-by-item ratings, any direct impact of providing such feedback will be greatly diminished by the process of item ratings. But, it does provide the opportunity for panelists to make adjustments to ratings during Round 3, and the size of those adjustments can be analyzed to determine the relative impact of this information on ratings.

The plan is to provide no feedback data to the other half of the panelists until the final round of ratings is completed. At that time, all panelists will be given information regarding the approximate percentage of students that will be classified at or above each achievement level, based on the final round of ratings. They will be

asked to comment on the differences between these percentages and their individual percentage estimates, as well as the percentage estimates averaged for the grade level. In particular, panelists will be asked to indicate which is/are more appropriate to recommend to NAGB and to comment on why the differences are large/small, important/unimportant, a result of the item rating methodology, a direct reflection of the lack of student knowledge and skills/a direct reflection of the lack of student motivation, and so forth. (Evaluation instruments will be developed for this purpose.)

b. Score Estimates

A similar question that proved troublesome in evaluating the 1992 data was whether the overall percent correct computed from each panelist's ratings at each achievement level corresponded to the panelist's notion of the score a student at each achievement level would get on a test over those items.

Panelists were shown a scaled average for their ratings at each achievement level relative to those of other panelists at their grade level. This information was given as interjudge consistency data. In addition, panelists were given intrajudge consistency data, after the second round of ratings, to inform them about those items for which their ratings were most deviant from their own individual "standard."

The plan is to give panelists this interjudge and intrajudge information again in the 1994 ALS processes. In addition, the pilot studies will be used to collect data on the effect of providing more information to the panelists. In this case, panelists will simply be asked to estimate the total score expected for a student meeting the minimum requirements to be classified at each achievement level if the student were administered a test with the items each panelist just rated.

Panelists in each rating group will be told the maximum score possible on their set of items. They will also be told that this score is based on dichotomous items being counted as 1 point each and polytomous items being counted as equal to the number of score values (e.g., "4" if the extended response item were scored 1-4, or "6" if it were scored 1-6. The number of score points assigned to the extended response items in the 1994 NAEP administrations is not uniform across extended response items within a single assessment.) The method of arriving at the total score (i.e., assigning points to multiple choice, short answer and extended constructed response items) will be described. The minimum score, based on the sum of c-parameters, will also be provided.

In order to minimize the cognitive complexity required to estimate a single score over several (60-80) items employing three different formats, panelists will be asked to estimate the score on each block of items rated. That is, after rating a block of items, panelists will be asked to estimate the score students at the lower borderline of each achievement level would earn on that block. Software will be developed to compute the average score over all items in the Item Rating Pool.

This information will be compared to the average percent correct computed from the item-by-item ratings for each panelist to determine how ratings of student performance, judged one item at a time, compare with the panelist's overall concept of how students at the margin of each achievement level would perform on the items in aggregate.

Again, items will be constructed for the evaluation instrument administered after Round 3 to get comments from panelists regarding the differences in scores estimated for student performance based on a single, aggregate estimate over a set of items *versus* aggregated score estimates computed from item-by-item estimates of student performance. They will be asked to evaluate the extent to which this information influenced their ratings for specific items, for specific achievement level(s), and so forth.

The plan is to collect achievement level score estimates from *all* panelists during each round of ratings. Only half of the panelists, the same half participating in the other "consequences" task, will be given the feedback data before ratings for *Round 2*. After the third round of ratings, however, all panelists will be presented with this score information. Panelists will be given information on the direct estimate scores and computed estimate scores for each round of ratings, and all will be asked to comment on the differences in their direct score estimates and their computed score estimates. In particular, panelists will be asked to evaluate the utility and relevance of this information for their ratings, the extent to which this information contributed/might have contributed to their confidence in their ratings, and the extent to which it contributed/might have contributed to the credibility of the achievement levels resulting from the process.

Whether and how these estimates will be incorporated into the ALS process will be determined after evaluating the pilot study results from the two approaches.

c. *Test Booklet Performance Data*

The question of how item-by-item ratings compare to more holistic ratings of student performance was troublesome to the evaluators for the National Academy of Education, as well. They recommended that the item-by-item rating methodology be abandoned in favor of a holistic approach. In particular, they recommended a "whole booklet" method as one of several to consider in setting achievement levels.

ACT had initially proposed a type of whole booklet methodology to be used as part of the validation studies. This "Whole Booklet" plan was described to collect information on the face validity of the achievement levels. A modification of that plan will be tested in the pilot study as a means of providing more information to panelists about student performance and to have that information portray a more comprehensive picture of student performance at each achievement level. The "score estimation" task in the previous section on "consequences" helps focus panelists' attention on the overall performance of students at each achievement level. The task described here sharpens that focus, perhaps, by giving information related to a test

booklet and performance of students on a set of items in the exact format administered to a sample of students who took the NAEP.

Early in the ALS process, panelists are administered form of the NAEP for their grade and content. Each of the NAEP booklets to be used for this exercise will contain two blocks of items to be completed in 25 minutes each. Panelists in Group A will be administered one booklet with items that will later be included in the item pool for rating by Group B, and *vice versa*.

As part of the feedback and information session preceding Rounds 2 and 3, panelists will be given a copy of the test booklet form administered to them earlier. They will be given the booklet score information for students at the grade level cutpoint for each achievement level. For example, students at the borderline of the Basic level would get an average of 37% of those items correct; students at the borderline of the Proficient level would get an average of 57% of those items correct; and students at the borderline of the Advanced level would get an average of 93% of those items correct, *given the level at which they have currently set their cutpoint*.[3]

Before rating items for Round 2, panelists will participate in a review of their achievement levels descriptions. This exercise, along with the additional exposure to the items in their Rating Pool, will help in that review. Panelists will be led to review their ratings and achievement levels descriptions in light of this booklet score data. They will be instructed to think about the items carefully and to determine whether that score seems about right for those items, given the students they have conceptualized as being characterized by the achievement level descriptions. If they believe they are about right, then their ratings are on target. If they believe that they are not right, then they can adjust their ratings during Rounds 2 and 3. They can adjust the achievement levels descriptions before Round 2, provided the adjustments meet with group approval.

All panelists will be provided p-value data (% correct, estimated *via* the three-parameter IRT model) for each item after Round 1. They will also be provided with inter-judge consistency data/graphs after Round 1 (and again after Round 2). They will have that data, the booklet scores, and the review/evaluation of achievement levels descriptions to help determine how they should adjust their ratings, if they feel ratings should be adjusted.

Again, items will be included in the evaluation instruments administered after this task is implemented. Panelists will be asked to comment on the utility of this

[3] The decision to present the data as performance at the borderline of each achievement level, as opposed to across the level, was made in order to minimize the complexity of the rating task and the potential for confusing panelists. Achievement levels describe a domain or range of performance, but the ratings (Round 1 of which will have just been completed) are of performance at the lower borderline of that domain.

exercise and on the effect of this exercise on their ratings, on their confidence in their ratings, and so forth.

d. Evaluations of Achievement Level Descriptions and Student Performance
Another suggestion that had initially been proposed by ACT for use in the validation studies will be implemented during the pilot studies to determine whether it can feasibly be included as a part of the ALS process, *per se*. This exercise will facilitate the final opportunity to refine the achievement level descriptions before Round 3, if it can successfully be added to the ALS process. It will require an estimated four hours to complete this task, however.

This exercise, referred to hereafter as the "Item Mapping Exercise," will address the question of whether the specific statements in the achievement levels descriptions are supported by performance on specific items in the NAEP pool for students performing within the ranges of the achievement levels set at that point, i.e., the cutpoints set at Round 2.

When panelists first start working with achievement levels descriptions at the beginning of the ALS process, they will be told that the achievement levels descriptions should be revised and refined to help ensure that they clearly define the skills and knowledge intended. During this exercise, to be implemented between Rounds 2 and 3, panelists will be given sets with all items categorized for each achievement level as follows. Note: **all** items will be classified for each of the three achievement levels according to the rubric described here.

1. The "can do" category will include items that have at least a .50 probability of correct response (for example, the probability could be set higher) at the *lower* bound of the achievement level. This will ensure that the probability of correct response to those items for students within the achievement level category will be greater than .50. It also ensures that students who score at any point higher than the lower bound of that achievement level will have a higher than .50 probability of correct response to those items.

2. The "can't do" category will be composed of those items that have less than a .50 probability of response at the *upper* bound of the achievement level category. These items will have less than a probability of .5 for a correct response for *all* students in that achievement level—even those at the upper bound. *Note: Because there is no upper bound for the Advanced level, only those items for which the probability of correct response at the lower bound is not greater than .50 will be excluded.*

3. The remaining items ("some can sometimes do") are in neither of the previous two categories. These residual items have the greatest potential to serve as exemplary of what students at the achievement level "should do." The probability of correct response to these items is ≥ .50 at *some score point(s) within* the achievement level range. Students scoring at the next higher

achievement level(s) have ≥ .50 probabilities of correct response. Some students within the achievement level in question and all students at the lower level(s) have ≤ .50 probability of correct response. This category thus includes *some* items that *some* students within this achievement level in question *can do* and *some* items that *some* students within this achievement level *can't do*. These items can be targeted to compare to the achievement levels descriptions which do, in fact, cover a range of knowledge and skills that students *should know and be able to do*.

The classification procedures described above will be used for dichotomously scored items. Polytomously scored items will require that the classification be made on the basis of student performance *at or above* each score category rather than for the performance task as a whole. (Because a partial-credit scoring model is used for these items, it is necessary to classify response scores as "1 or higher," "2 or higher," and "3 or higher.") For the same performance task, a score of "3 or higher" on a 5-point scale may have a >.5 probability for student performance at the lower bound of the Proficient Level, while a score of "4 or higher" may have a probability below .5 at the upper bound of that level.

Panelists will examine the two categories of "can do" items: those that all students have a ≥ .50 probability of answering correctly and those that students at some point(s) within that level have at least a .50 probability of answering correctly.

Panelists will examine items in these categories at each achievement level to determine whether those skills correspond to the statements included in the descriptions. Similarly, panelists will examine the "can't do" items for each achievement level to determine whether descriptive statements are found to lack confirmation by student performance on the items. Some statements will perhaps be found for which the item analysis will provide confirmation or the lack thereof. This can result from there being no items on the NAEP to assess the skills or knowledge described or because the item level results are simply too ambiguous.

Panelists will be given time to modify achievement levels descriptions within grade level sessions. These modifications will be made with the assistance and guidance of content experts on site to assist panelists in tasks related to expanding and refining achievement levels descriptions.

This task can be considered in the "consequences" category because it will focus panelists' attention directly on the items that students "can do" and "cannot do" and on how those specific items relate to the achievement levels descriptions used to rate the items and classify them as such. Devoting time to this task before the final round of ratings helps ensure that ratings are based on a common understanding of a common set of descriptions used by all panelists at each grade level. By the final round of ratings, **all** panelists will have the following information:

1. achievement levels descriptions: expanded and refined before each rating period (extensively before Round 1, before Round 2 with stimulus of "Whole Booklet" exercise (which will be updated and distributed again before Round 3), and before Round 3 with stimulus of "Item Mapping" exercise)

2. student performance data: estimated p-values for dichotomous items and polytomous items (before Round 2)

3. interjudge consistency data (before Rounds 2 and 3)

4. intrajudge consistency data (before Round 3)

5. performance by students at each achievement level on a NAEP test booklet ("Whole Booklet" exercise before Round 2 and updated before and after Round 3), and

6. "Item Mapping" data to classify items that students at each achievement level can/cannot do (before Round 3)

In addition, half of the panelists will be given feedback before the final round of ratings to inform them about the percentage of students who would score at or above each achievement level set in Round 2, the mean percentage correct resulting from the item ratings to set the achievement levels at Round 2, and the mean "test score" estimated for the items by the panelists (given as feedback before Round 2).

▪ ***Research on Item Rating Methodology for Polytomous Items***

The results of the pilot study for the 1992 ALS process revealed that ratings for polytomously scored items, i.e., extended response items, led to higher achievement levels than ratings for dichotomously scored items, i.e., multiple choice items and items that were scored as either correct or not. Some additional data were collected for research purposes during the ALS processes implemented, but those could not be used to arrive at any conclusive explanations for the differences.

ACT feels that it is extremely important to the future of standard setting with extended response and performance items that alternative rating methods be tried out. The pilot studies provide the opportunity for trying out alternative rating methods that appear to be conceptually concise and technically sound. The results will be evaluated to determine which to incorporate into the ALS process. Criteria to be considered in the evaluation of methodologies include the evaluations from panelists regarding the level of confidence in ratings associated with each method, the level of clarity and certainty each panelist reports regarding the cognitive task involved in applying the methodology, and the amount of time required to implement the training and ratings for each method. In addition, quantitative analyses of the level of variability between ratings for polytomous and dichotomous items will be conducted to compare the rating methodologies. Moreover, ratings using the

alternative polytomous methodologies will be compared with respect to the interjudge consistency and intrajudge consistency, for example.

Four different methods for rating polytomous items will be tried out during the pilot studies: two methods will be tried for both geography and U.S. history, and a third method will be tried in each. The paper selection method used in the 1992 ALS process will not be implemented as a rating method, *per se*, in the 1994 ALS process. Rather, this methodology will be used with all panelists during the training and preparation for item ratings. In addition, one method to be tried out in the pilot studies directly incorporates the paper selection methodology into the procedure for the first round of ratings.

Each method will be used for rating the polytomously scored items by one Item Rating Group at two different grade levels. For both geography and U.S. history pilot studies, Groups 4A and 8B will estimate the percentages at each achievement level for each score point, and Groups 8A and 12B will estimate the mean score for each achievement level. For the geography pilot study, Groups 4B and 12A will use the estimated mean scores method, and those groups will use the modified method of percentage estimates in the history pilot study.

1. Estimated Mean Scores

One method expected to be the easy for training and implementing simply requires panelists to estimate the mean scores for students at the lower borderline of each achievement level. If the task is scored on a three-point scale, for example, a panelist might estimate that the mean score for borderline Basic students would be .7, while the mean score expected for borderline Proficient students would be 1.9 and that for borderline Advanced would be 2.7. We anticipate that one decimal point will provide adequate specificity for panelists, but up to two digits can be coded.

This method will be implemented in the geography pilot study.

2. Estimated Score Point Percentages

The second method to be tested during the pilot studies requires that the panelists estimate the percentage of borderline students at each achievement level who would be scored at each point on the score scale for the extended constructed response (polytomously scored) item. This methodology has been examined *via* simulated ratings and found to produce results that scale to points that are not significantly different from results obtained for ratings of dichotomous items rated *via* a modified Angoff procedure.

This methodology is conceptually the same as that to be used for rating all dichotomously scored items, and it is expected that panelists will not require extensive, special training for rating the polytomous items with this method.

A potential difficulty with using this method can be eliminated rather simply. Early reviewers of this proposed method suggested that panelist would find it too hard to

use this method if they were expected to estimate percentages that summed to 100. That is, if panelists estimate the percentage of borderline Basic students who would score "1," the percentage who would score "2," and so forth, the sum should be 100% for borderline Basic students. That *could* be quite burdensome for panelists to have to adjust ratings for each score point so that the sum is 100%. Software will be developed to recompute, with a base of 100, each set of achievement level estimates to sum accurately.

3. The Modified Percentage Estimate

A method that modifies the previously described method will be tried out in the U.S. history pilot study. A score of "1" is assigned to student responses that are generally described as lacking in some way. Thus, only a score of "2" or higher is deemed to be addressing some of the essential points needed for responding to the task. This rating method requires panelists to estimate the percentage of students who would provide a response that would be scored as "2 or higher."

4. The Hybrid Method

ACT had initially proposed using the paper selection method again, as one of two methods proposed, in order to collect data to use in investigating the differences revealed in the 1992 ALS processes in ratings between polytomous and dichotomous items. The TACSS recommended that priority be given to trying out the alternative first two procedures identified above. In addition, however, they recommended that third method be tested in the pilot study of each content area. The modified percentage estimation described above will be tested as the third method in the history pilot study, and this Hybrid Method will be tested in the pilot for geography. The hybrid method is a combination of the paper selection method and the mean score estimate.

Panelists will be given a maximum of thirty samples of student's responses to performance-type tasks. The scores for these samples will not be given to panelists. The papers to be used in this process will be the papers used by National Computer Systems (NCS, the contractor to ETS for NAEP scoring) in training scorers. These calibration papers may be supplemented with additional papers used in the training process in order to obtain the needed distribution of papers. The plan is to use a rectangular distribution of paper scores. This means that, to the extent possible, an equal number of papers will be included for each score point for the task. If there are six score points, the goal will be to include five samples of each. If only three students were scored at six, for example, then the goal will be to include seven samples of responses scored at five. (The alternative to the equal distribution of papers at each score point would be to present a representative sample of paper scores. This presents the real possibility that all but very few papers, perhaps only one or two, would be scored at only one or two score points. Panelists would have no possibility of selecting a paper to represent performance of students in all three achievement levels. Thus, we have opted to employ the equal distribution design.)

Panelists will be asked to select up to three papers to represent student performance at the lower borderline of each achievement level. The complete scoring rubrics will be provided to the panelists to help guide the selections, and they will use the achievement levels descriptions as well. Panelists will be asked to estimate the score for each paper selected.

That methodology will be used for the first round of ratings. When the Round 1 results are presented to those panelists, they will be told the scores actually assigned to the papers they selected. In addition, all panelists at each grade level will be given the frequency distribution for student scores on the polytomous items. The panelists participating in this rating process will be told the mean score for the papers they selected in Round 1 to represent each achievement level. If they feel that these are appropriate, they may use those means as their ratings in the subsequent rounds. If they feel that these scores should be adjusted, they can record the mean score that they believe students at the borderline of each achievement level would earn on each extended constructed response task in their Item Rating Pool. In other words, ratings for Rounds 2 and 3 will be provided using the same methodology as that described in the first alternative above—the mean score estimates.

This hybrid method provides these panelists greater familiarity with actual responses that students give to the extended constructed response tasks. While all panelists will have training in the paper selection method in order to become familiar with the scoring rubrics and the range of student responses, this method will focus panelists on these samples as the first step in the rating process. However, because it is difficult for panelists to change their "ratings" using the paper selection methodology, they will have the opportunity of making adjustments *via* the mean score estimation procedure.

We learned during the 1992 process that panelists were somewhat frustrated by the fact that papers were apparently scored quite differently than they had expected. Panelists were never given the paper score, *per se*, but they were given the average over papers they selected and the range of points for those papers. This information was sufficient to indicate the scores of the papers selected, but not to identify a specific score with a specific paper. Information from panelists about the scoring will be helpful in understanding their ratings of these performance assessment items. Every effort will be made to train panelists in the scoring rubrics so that they are competent to judge the papers according to the appropriate criteria. Further, every effort will be made to keep panelists "calibrated" with respect to *the* rubrics so they do not resort to using their own rubrics. If panelists' evaluations of papers, represented by the scores they assign the papers, do not generally correspond to those recorded for the students, then the paper selection method should not be used. Guidelines must be established in advance to determine how much variance overall, for each individual panelist, or for individual items, will be considered tolerable for evaluating the feasibility of incorporating this method in the process for setting achievement levels with performance assessment items.

▪ ***Summary***

In summary, the pilot studies will provide the opportunity to implement and evaluate all aspects of the operational plan—background materials, meeting materials, study design, validation procedures, meeting logistics, staff function, and participant function—in a thorough, yet efficient, manner. Results of the pilot study will be fully reviewed and discussed with the TACSS, with NAGB staff, the Achievement Levels Committee, and the Board to ensure successful implementation of the procedures in the actual process. Specified groups, such as members of TACSS, representatives of the content and technical staff from ETS, representatives of NCES, representatives of the evaluation team (if there is one) could be invited to observe the pilot studies. They will be asked to provide information and feedback for our review and evaluation in order to maximally improve the process and information collected.

Finally, the results of the pilot study will be distributed to stakeholder groups for review and comment. That review period is tentatively scheduled for September 12-16, 1994 for results of the Geography Pilot Study, and October 3-7, 1994 for results of the U.S. History Pilot Study. The Science Pilot Study results will be distributed to stakeholders for review June 10-14, 1996, according to the current, very tentative schedule for science.

Table 2 presents the various procedures to be implemented in the pilot studies. This table provides a quick overview of the process and the groups involved at each stage. A "draft" agenda for the pilot studies is provided in Figure 3 to provide some idea of how the many tasks to be performed and data analyses to inform the tasks during the pilot studies can be worked into a five-day schedule.

Implementation

Several important sets of outcomes are expected from the achievement levels-setting meetings. The primary outcomes are the content-based descriptions of each of three levels of achievement for each of the three grades; numerical cutpoints, on the theta metric, defining the lower bound at each achievement level; and items, available from the 1994 pool of items slated for public release, for selection as illustrative of the skills and knowledge characterizing student performance at each of the three achievement levels at each of the three grades assessed in each content area. Other outcomes are less tangible, but very important from the perspective of the credibility and eventual utility of the resulting levels and their descriptions. Some of these include the information on conceptualizations of achievement, reactions to data on consequences of ratings, and panelists' evaluations of student performance relative to expectations gathered during the pilot studies and ALS implementation meetings. Others are represented by the confidence and satisfaction of the participants with respect to the process, the resulting numerical levels, and their content-based descriptions.

A five-day meeting is planned. Additional activities have been included in the process for 1994 that were not included in the 1992 process. Some time has been

added to the agenda scheduled for the beginning and ending days. While more time would be desirable, ACT deemed it injudicious to ask panelists to devote more than five days of their time to this effort or to expect that panelists *could* actually give more than five days to activities of this intensity. We must, however, ensure that sufficient time is available for every key task in the process. We must eliminate all possible sources of confusion and dissatisfaction on the part of the panelists with respect to what is to be done, the purpose for doing it, and how it is to be done.

- Achievement levels panelists must attain fluency in the assessment framework and initial descriptions in order to successfully expand and refine these descriptions before beginning the item rating task.

- Panelists must become proficient in the item rating methodology, and they must have enough time to complete the item rating tasks with care and attention.

- Panelists must be trained to carry out the exercise to compare achievement level descriptions with items classified according to whether students at an achievement level can do/cannot do them.

- Panelists must be trained in the use of all feedback information to be used in rating items.

- Item ratings must be input into a database and the accuracy of data entry must be verified. All analyses needed to provide the feedback called for in the design of this process must be performed with the highest level of accuracy and efficiency.

- Evaluation data must be collected from panelists each day and after each major task or activity.

▪ ***Orientation to ALS Meetings***

Panelists are asked to spend five days working with our staff to carry out the process described in this document to set achievement levels on one of three different NAEP content areas. These panelists come from diverse backgrounds and experiences. One of the first activities planned for the panelists is a social hour when panelists and staff can begin to get acquainted. This opportunity is particularly beneficial because so much time is spent in grade level groups that it becomes difficult to become acquainted with panelists and staff from other grade groups. This time also seems to give some assurance to panelists that (s)he is not the lone novitiate to the process.

The **first element** of the Achievement Levels-Setting (ALS) process focuses on providing a common understanding of the purpose of setting achievement levels and the procedure to be followed in setting the levels. Panelists must know what is expected of them and how they will be helped to achieve those expectations. ACT proposes to give panelists information about the purposes of the assessments, the

purposes of the achievement levels, and the purposes of all information presented to them and collected from them. This dispels doubts and fears of being "used" for purposes that are not known and, therefore, suspect. The first session is critical to the success of the entire process, and it will be attended by the whole group (that is, by the 30 fourth-grade, 30 eighth-grade, and 30 twelfth-grade subject area panelists). Concerns and misgivings must be eliminated as early and as completely as possible. This means that they must be addressed fully and candidly at the very beginning of the process.

It is essential that meeting participants be familiar with the purposes for setting NAEP achievement levels and that they have a general understanding of the roles of the principal participants in the process. To this end, it is proposed that the initial session be jointly conducted by the ACT project director and a representative of NAGB (the COTR, for example) who will serve as an expert resource on the background of NAEP and NAGB and the developments leading to the specific assessment for which achievement levels are to be set.

The first session begins with a welcome and general orientation, a description of the meeting agenda and the tasks to be accomplished, and an explanation of how panelists were nominated and selected. The session will include an overview of NAEP and changes in the assessments over time, an overview of NAGB and their role in NAEP, a presentation concerning NAGB's policy (generic) definitions of achievement levels and the initial definitions for the specific content area, and an overview of the process to be implemented during the next four days.

At the end of the first general session, incorrect and unwarranted assumptions are eliminated and replaced with information about the procedures to be followed and the purpose for doing so.

Much of the work of the panelists during the ALS process takes place in the grade level groups. A facilitator will lead the activities of each grade level group, and it is clearly necessary that the facilitator reinforce the sense of confidence and integrity initiated during the opening general session. The facilitator assigned to each group will possess several key skills: training and experience in conducting judgmental standard-setting procedures; strong interpersonal skills; and an ability to discern group dynamics to foster efficient, focused activity. In addition to the three facilitators, one content person will be assigned to each grade panel to serve as a resource person and provide content area input as requested by the groups. The ACT Project Director, who has extensive experience in managing group dynamics, will constantly monitor group interactions to verify that the group process is functioning effectively. The Project Director served as a facilitator, coordinated the facilitators, and led in detailing the plan for each session in the 1992 ALS process.

Also a part of creating a comprehensive understanding of purposes and means of achieving them is the more in-depth, detailed description of the ALS process that will be provided at the start of Day 2. The computerized flow chart description to be

developed of the process will ensure consistency of the information presented and it will add interest by providing information through another medium. The data show will supplement, not replace, presentations by speakers. The central point is the fact that the flow chart description of the process will be presented throughout the five days of the ALS process. By revisiting the information again and again, panelists will more completely internalize even the more subtle features of the tasks being accomplished. They will have a clear conception of what they have accomplished and what remains to be accomplished.

■ ***Expanding and Refining Achievement Levels Descriptions***

The **second element** in the ALS process is the expansion and refinement of operational descriptions of the achievement levels at each grade. For the first time, the framework for the 1994 NAEP assessments include descriptions of achievement levels specific to the content area for each grade—4th, 8th, and 12th. These initial descriptions provide firmer grounding to the panelists who must become conversant with these descriptions. In fact, one of the most important elements in a successful standard-setting process is reaching a common understanding among raters on the definition of the achievement levels. It is essential that panelists have a clear and common definition of what students should know and be able to do at each achievement level for their grade in the specific content domain. Without this common agreement on the meaning of the achievement levels, the ratings have no interpretation beyond the numerical values that are identified.

Meeting participants must be thoroughly familiar with the definitions of NAEP achievement levels before any item rating tasks are performed. The initial definition of the achievement levels as specified in the content framework will be presented along with a detailed description of the conceptualization and philosophical foundations of the assessment framework. The plan is to provide the foundation for more in-depth development and operationalization of these achievement levels. The process for accomplishing this will incorporate general sessions, grade level sessions, and within-grade work and discussion groups. A variety of activities will be performed during the implementation of the part of the process. Over one full day in the five-day process is devoted to arriving at a common understanding of each achievement level and to having those definitions form a logical progression within each grade across achievement levels, as well as across grades within each achievement level.

a. Training in Framework

A whole-group meeting during the beginning of the ALS process will include a presentation on the Assessment Framework for the specific content area. ACT will ask for assistance from NAGB staff, representatives of the framework consensus panels, and representatives of key professional organizations in the content area to identify an outstanding person to make the presentation for this session. The goal will be to identify a person who is very familiar with both the content and development of the assessment framework and who is also very good at public speaking. The person selected for this role must not only inform the panelists about

the framework and initial achievement levels, but that person must also present this information in a manner that will contribute to the sense of confidence in and integrity of the process. Careful screening of potential presenters is essential to ensure that everyone involved in the process is focused on the same goals.

b. Experience with NAEP Items and Tasks

One of the first steps in preparing panelists to refine and expand the achievement levels descriptions will be the opportunity for each panelists to gain experience with an actual form of the NAEP subject area assessment for their grade level. ACT concurs with NAGB guidelines that panelists must be familiar with the content of the assessment that they will be rating. Accordingly, prior to providing ratings, each panelist will take one form of their grade-subject area assessment under timed conditions similar those experienced by students. At each grade level, one booklet form will be administered to half the panelists and another to the other half. The item blocks included in the booklets used in this administration will not be included in the item rating pool for the panelists. (See the description of the test booklet performance data exercise in the Pilot Study section.) Panelists will be given scoring keys and protocols to use in scoring their own examinations to facilitate their understanding of the items, "correct" answers, and scoring methodology.

This element is proposed so that panelists are familiar with the general content covered by the assessment, the time constraints imposed, and the general level of difficulty of the assessment. Panelists can then begin crystallizing their conceptions of the three achievement levels with respect to both the descriptions they are developing and the actual test content.

Strict security arrangements will be followed to ensure that all materials are accounted for at all times.

c. Agreement on Operational Descriptions

Most of the work on expanding and refining the achievement levels descriptions will take place in grade-level sessions. The work in these sessions will be completed in six units (table groups) of five panelists for each grade group. Three tables, with five panelists each, will be rating Group A and three will be rating Group B. Work groups will gradually be increased in size, by combining table groups, so that the size of the group working on refining and expanding descriptions is larger and larger and the size of the group reaching agreement on these modifications is larger and larger.

Facilitators and content specialists will lead the panelists in exercises to help them internalize the descriptions in relation to the frameworks, to identify holes or gaps in the descriptions or their meaning relative to the frameworks, and to become comfortable and conversant with the descriptions.

As a starting point in arriving at agreement on the achievement levels descriptions, panelists will be asked to work in groups of two or three, and to start verbally expressing and paraphrasing their individual understandings of the initial

achievement levels descriptions. The small interaction teams will be asked to discuss their ideas with others in their table group, with others in their Item Rating Group, and then with others in the entire grade group. This will help establish targets for the grade group for their work on refining and expanding the definitions.

Another method to facilitate panelists in evaluating the descriptions and identifying needed modifications is to parse the draft descriptions and have the panelists evaluate key components (e.g., the adjectives and verbs) of the sentence segments or clauses independent of the supporting sentences or paragraph. The evaluation would be of the segments within each level and of segments across levels. This exercise would be beneficial to panelists after they have worked with the initial definitions long enough to have developed a good understanding of their meaning. This exercise is aimed at delineating the redundancy, ambiguity, details, and jargon. Panelists can negotiate among themselves, and with the assistance of the content specialists, whether these identified features should be changed.

As part of the process of reaching agreement on the achievement levels descriptions, panelists will review items that they will **not** be rating later. One such exercise involves practice in using the achievement levels descriptions. Panelists will be given sets of 10 or 12 items and asked to evaluate them according to how they think students in each achievement level would perform on those items. Panelists might be instructed to identify items that they believe *most* (e.g., at least two-thirds) of the students at the Basic level of achievement would get correct; those that *few* (e.g., less than one-third) of the students at that level would get correct; and those that *some* (e.g., from one- to two-thirds) of the students at that level would get correct. Panelists would be asked to consider these items for the Basic, Proficient, Advanced and "Below Basic" levels. Items included in the rating pool for Group A can be used by Group B, and *vice versa*. This will help panelists gain a more complete sense of the full array of items included in the grade-level pool. The evaluations will be done independently, but panelists will discuss their evaluations in table groups or rating groups in order to gain an understanding of how members of the group differ in their conceptualizations of the levels. This exercise will further aid in identifying strong and weak features of the descriptions. Following this discussion, panelists will work to further refine and expand their descriptions.

Throughout the process of refining the achievement levels descriptions, panelists will be directed to refer back to the framework to determine whether the descriptions are within the parameters of skills and knowledge encompassed in the framework. Content consultants will be assigned to each grade group to help provide the interface between frameworks and descriptions.

An important aspect of the NAGB achievement levels is the fact that the descriptions are of what students *should* know and be able to do. The achievement levels define a range of knowledge and skills that form the domain of each level. Item ratings, however, must be targeted specifically at a point. Panelists will be asked to focus on the lower boundary of each achievement level—performance of students that

minimally qualify to be included in an achievement level—in order to provide their item ratings. Further, when rating the performance of the minimally qualified student at each achievement level, the panelist will be asked to think of how the student *will* perform. These distinctions are commonly made in standard setting as a device to focus attention on a common set of attributes and a specific point within that set. This requires that panelists have a very clear concept of the range of the achievement levels in order to target the lower bound of that range. Panelists will be made aware of the distinction between the domain of knowledge and skills represented by performance within the achievement levels and the boundary demarcating minimal performance at each level.

After panelists have spent an entire day working in the grade-level groups, a general session will be reconvened to share the results across the three grade groups. Content specialists will have been in communication with one another to monitor the developments across grade groups. They will have helped prevent any great departures on the part of any one grade group, and they will be prepared to point out significant differences that have developed during the day.

During the evening, content specialists will meet to discuss the definitions developed to that point by the three grade groups. A teleconference to include additional representatives of the framework panel or stakeholder groups can be scheduled as well. They will identify further refinements and modifications to offer as suggestions to the panelists. An additional grade-level work session is scheduled to provide panelists an opportunity to evaluate the recommendations of the content specialists and discuss the descriptions for their grade, relative to those of other grades. They will then spend time making further modifications to their descriptions and arriving at grade-level agreement on the descriptions to guide their ratings for the first round.

Panelists will be aware of the fact that they will have additional opportunities to make adjustments in the descriptions before the final round of ratings. They will also be aware, however, of the fact that those adjustments must be relatively minor if the results of each round of ratings are to be of value in guiding their subsequent ratings. Thus, a common understanding and agreement should be reached by panelists regarding the achievement levels descriptions before item ratings are undertaken.

While it is essential that panelists arrive at a common understanding of skills and knowledge encompassed by each achievement level description, it would be unwise to ignore the learning that will take place during the rating process. A general strategy of ACT to be employed throughout the ALS process is to provide iterative training sessions. After applying these descriptions in the item rating process, panelists will have a keener sense of description nuances and will identify additional modifications required for reaching closure on the process of expanding and refining the descriptions.

Tasks will be tried out during the pilot study that will potentially contribute significantly to the panelists' ability to reach closure on the task of expanding and

refining the descriptions. One such task will be carried out after two rounds of ratings have been completed and before the final round—the round used for estimating the recommended cutpoints—is undertaken. That task is referred to as the "item mapping" or "can do/can't do" exercise in the description of the Pilot Study in Section 2.

■ ***Training Panelists as Raters***

The **third element** in the ALS process is training in the rating process. Thorough training in the item rating methodologies is essential to the judgmental rating procedure (Francis and Holmes, 1983; Klein, 1984; Livingston and Zieky, 1982). ACT will rely on its own well-developed resources for instructing panelists in the methodology, as well as on the input of NAGB staff, the technical advisory groups (TAT and TACSS), and from other appropriate groups (e.g., the evaluation team report of the 1992 process prepared by the National Academy of Education).

The training will begin with review of the achievement levels descriptions developed in the earlier sessions and build upon the concept of minimally acceptable performance at the Basic, Proficient, and Advanced levels. Achievement level descriptions are of the performance of students within a level. Ratings, however, are made of performance at the lower bound of each achievement level.

Training for each subject area will be customized to reflect the unique configuration of item formats and performance tasks for the three subject area assessments. A general session will be convened to provide a common core of training to all panelists. During this session, panelists will see rating forms and will learn how to mark them. A few items of various formats will be used as examples to demonstrate how to rate items. Panelists will be instructed to rate each item at all three achievement levels before going to the next item. They will, however, be given the option of rating the achievement levels in any sequence they find most helpful: Basic, Proficient, Advanced; or Proficient, Basic Advanced; or Advanced, Proficient, Basic, and so forth.

Panelists will be asked to answer each item before rating it for the first round. They will also be instructed to refer to the scoring guides for the correct answer for each item before rating it. This will help panelists have a better sense of the difficulty of the item than simply reading it and rating it. Training in the use of feedback and other information to be provided to panelists for ratings in Rounds 2 and 3 will be provided prior to those actual rating rounds.

Following the general session, panelists will have further training in grade-level groups using items from the field trial pool, if they are of adequate quality, or actual NAEP items from the half of the item pool they will not be rating later. Ten dichotomously scored items and one or two extended response/performance tasks will be rated and reviewed for training. These items will represent, to the greatest extent possible, the full range of content, item type, and item difficulty in the actual pools to be rated.

Recognizing the complexity of the task, ACT has designed the preparation for this element specifically to allow ample time for panelists to learn the rating task, to ask questions, to receive individualized feedback, and to become comfortable with the rating process. It is important to keep in mind that the panelists are instructed to provide their ratings in the context of the skills and knowledge that a student at the lower bound of each level should possess, and then to estimate the proportion of these minimally performing examinees that would answer in the specified ("correct") manner. Thus, the rating reflects the panelists's estimate of the performance of students *just at* each of the three levels.

■ ***The Item Rating Process***

The **fourth element** is the actual rating of test items. ACT recommends that three rounds of ratings be collected. The rating process will occur within grade level groups, and the grade level groups will be further divided into subgroups of fifteen panelists. Assignments of panelists to the two groups will be done according to random assignment within strata. In the first round of ratings, grade level groups will provide ratings for each item pool without reference to item statistics. Panelists must answer each item and refer to the scoring guides to determine how answers were scored. This will require a considerably longer time for rating items than in subsequent rounds. Panelists will be asked to provide three ratings for each item—the proportion of examinees at each achievement level who will answer the item correctly.

Approximately five hours will be scheduled for the Round 1 ratings. If the *Hybrid Method* (including paper selections) of rating extended response and performance tasks is used, even more time may be required for panelists at Grade 12. This method will be assigned to Grades 4 and 12 during the pilot studies in order to determine the minimum and maximum amount of time needed for the task for each content area.

Items for each half of the item pool will remain in blocks and in the order they appear in the blocks. The item pool halves will be selected to be as equal as possible in terms of the overall student performance/item difficulty, item content, and according to item formats, and additional materials (cartoons, atlases, and so forth) accompanying each block of items. In addition, at least one block of items will be rated by all panelists within each grade level. These provisions allow ACT to analyze ratings for each item pool half as if they were actual replications of the ratings. Having a block of items rated by both groups allows a direct check on the effects of different raters.

In addition to the item ratings, all panelists will be asked to estimate the score that a student at the lower bound of each achievement level would get on the set of items just rated. The total possible score will be given to panelists, and the method of arriving at the total score (i.e., assigning points to multiple choice, short answer and extended constructed response items) will be described. The minimum score, based on the sum of c-parameters, will also be provided. Further, all panelists will be asked

to estimate, for the grade level of students, the percentage of students who would score at or above each achievement level. These two procedures will be tried out during the pilot study with a treatment group and a control group in order to determine whether either or both can be used during the ALS process and how each should be used, i.e., to inform panelists or as validation information for the process and outcomes.

The rating procedure is based on collection and analysis of ratings of individual test items by knowledgeable, qualified judges or panelists. In a typical implementation of the procedure, content experts are asked first to identify the skills, knowledge, and performance characteristics of the minimally acceptable candidate for a particular classification (e.g., pass/fail, novice/expert). Once that characterization has been developed and agreed upon, the panelists are then asked to estimate, for each item, the proportion of a group of 100 such minimally acceptable examinees that *would* respond correctly on each item. This procedure is typically applied to multiple-choice items. Once the panelists have estimated, for each item, the proportion of the group of minimally acceptable examinees that would answer the item correctly, those proportions are averaged across items and panelists. This average is the proportion of items that must be answered correctly by an examinee in order to be classified as minimally acceptable for that particular classification or title.

For *this* project, the procedure must be modified to accommodate items that are not scored as right or wrong/correct or incorrect and for which score distributions are provided, as opposed to p-values (i.e., the percentage of correct responses). For the multiple-choice and short answer items—those scored dichotomously—the item rating procedure is to be directly applied. The panelist's ratings will be converted to a theta value for each item. Considering each item as a replication, the distribution of theta values will be determined and the mean identified as the boundary estimate for that panelist. The distribution of boundary estimates will be determined across panelists for each level, and the mean of that distribution will be identified as the numerical lower bound for the given achievement level and grade.

For the extended response item types, the panelists will be given the stimulus prompt and scoring protocol along with a sample of actual examinees' papers. The procedures for rating extended response items were described for the pilot studies in Section 2. The requirements of two of the three alternatives to be tried in the pilot studies are only slightly different from those for the basic item rating process used for dichotomously scored items.

If the Hybrid Method (paper selection as Round 1 with mean score estimates for subsequent rounds) is implemented in the ALS process, panelists will select up to three papers that exemplify the performance of minimally acceptable examinees at each achievement level for their Round 1 ratings. The score for each paper selected will be converted to the theta scale, then aggregated across items and panelists just as for the dichotomous items.

All panelists will participate in a paper selection process during the general training period or during the period of working with the achievement levels descriptions. This exposure to actual student responses to the extended constructed response tasks seems necessary, regardless of the item rating methodology implemented to set achievement levels.

Panelists will complete the first round of ratings with no information except the definitions worked out in the previous day for the three levels of achievement (i.e., Basic, Proficient, and Advanced). For Rounds 2 and 3, panelists may change their percentage estimates or mean score estimates (if that methodology is used for rating polytomous items), or they may keep their ratings from the previous round. This is true for each item and for each achievement level. For example, a panelist may wish to adjust ratings at one or all achievement levels for a specific item or set of items or for a single achievement for all items, and so forth. A change in any rating is permitted *during* each round, but ratings *for* a previous round are not.

■ ***Informing the Ratings via Feedback and Other Information***

The **fifth element** in the ALS process to be implemented is the provision of information and feedback. Before each subsequent round of ratings, a general session of panelists will be convened. At this meeting, panelists will be provided with a brief retraining in the rating task. They will also be given information regarding the average ratings for each achievement level at each grade. Discrete NAEP items will have been calibrated using a three-parameter logistic IRT model (Johnson & Mislevy, 1991; Lord, 1980). These parameters will be available from Educational Testing Service (ETS), the current NAEP Operations Contractor. The data will be scaled and the estimated item characteristics used to determine the latent scores and other feedback information for panelists. For both the dichotomously and polytomously scored items, the estimates of the latent traits will be obtained immediately following each rating using software developed by ACT specifically for this project. To help the panelists interpret these latent trait estimates, ACT has developed software that will report the estimates as proportions of the entire NAEP item pool. Graphic presentations of individual and group distributions of estimates will be relied on extensively for this presentation to assist panelists in understanding the results.

Because the 1994 NAEP data are to be reported on a scale that is computed within each grade, as opposed to across grades, the discussion will focus on the achievement levels within each grade. Items that are assessed at more than one grade can be evaluated and the achievement levels, based on those items alone, can be evaluated for consistency across grades.

The primary focus of the meeting, however, will be to orient the panelists regarding information they will receive for their next round of ratings. Panelists will be given extensive feedback to inform them about their ratings, and they will be provided data in manageable quantities. The information is to further aid their conceptualizations of how students at each achievement level would perform on each item.

a. Information targeted at improving "intrajudge consistency"
To provide this feedback, ACT will utilize an item response theory (IRT) model to determine a transformed NAEP "score" for each panelist.[4]

For example, for the discrete 4th grade geography items, panelists will estimate the proportion of examinees who minimally qualify at the lowest level of the Proficient category that will respond correctly to each item. That proportion can be used, along with the previously obtained item parameter estimates for the item, to estimate a latent trait score that defines the lower bound of the Proficient category. A separate estimate of actual student performance can be computed with respect to the lower bound cut score and compared to the panelists' ratings or paper selections. If these comparisons are very similar to each other (i.e., have small absolute deviations), a panelist will be deemed to be very "consistent" in judging the items. If the estimates of actual student performance (conditioned on the standard) and panelists item ratings or paper selections vary greatly, the panelist is being "inconsistent," and the item ratings that are the most deviant can be identified. By identifying these most deviant items for the panelist, it should be possible for the panelist to determine some reason for the inconsistency, and this should lead to more consistent estimates in subsequent rounds of ratings (Luecht, 1993). All feedback data will be for the consideration of panelists to facilitate their ratings in subsequent rounds. Staff will stress the fact that the data are to inform panelists—not to coerce their ratings decisions.

b. Information targeted at reducing "interjudge variability"
Reduction of interjudge variability will be accomplished through the provision of item-level data on the actual performance of students on the test items. These data will be in the form of item difficulty values for each item. For dichotomously scored items (those scored correct/not), this information will be the percentage of students who correctly answer each item. For extended, or graded, response items, the information will be the percentage of students scoring at each point on the score scale. For all item types, the percentage of students who omitted or did not answer each item will also be reported.

In addition, the distributions of the individual panelist's cutpoint estimates can be used to compare the ratings among the different panelists. The panelists that are most extreme in their ratings can be identified. Their ratings can be discussed to determine whether they have a different understanding of the tasks required by the items or by the meaning of the achievement level descriptions. Graphics have been developed that help panelists see where their standard would fall relative to the

[4] The linear transformations used for the 1992 assessments placed each panelist on a relative scale having a mean of 75 and standard deviation of 15. These transformations scaled the panelists' standards to be between 0 and 120. This relative scale facilitated interpreting ratings and feedback data for the panelists. A similar transformation will be used for the 1994 assessments.

mean for their grade and relative to every other panelist at that grade level. Previous research has recommended such procedures and has indicated that they are often effective in reducing unwanted variation and in helping to promote consistency and convergence in ratings (cf, Berk, 1986; Conaway, 1979; Livingston & Zieky, 1982, and Shepard, 1979; 1980a; 1980b; 1983). In addition to these data, panelists will also be repeatedly referred to their common referents (e.g., the group-produced definitions) to further promote consistency and convergence during the second and third rounds of ratings.

c. Information aimed at informing panelists about their ratings

ACT will study the effect of collecting and presenting information targeted at anchoring the ratings in *reality*. The "consequences" tasks are described for the pilot studies in Section 2. ACT will collect information during the pilot studies to evaluate the effect of information about the empirical distributions of student performance with respect to achievement level cutpoints, and about judgements of student performance based on sets of items taken as a whole rather than item-by-item judgements. The results of the pilot studies will be carefully reviewed by the TAT and TACSS before making a recommendation to NAGB regarding the final design to be implemented as the ALS process in each content area.

In addition, ACT will implement procedures in the pilot studies for evaluating the types of skills and knowledge that students at each achievement can and cannot do, based on achievement levels set in ratings just prior to that evaluation. The items will be evaluated relative to the achievement levels descriptions to determine whether the panelists perceive the levels to be appropriate, in light of the information about how students performed on the items in their rating pools. If this procedure can be successfully implemented and if the results indicate that the procedure is useful in informing the panelists about the relationship between their item judgements and the achievement levels descriptions, the procedure will be incorporated into the ALS process, *per se*.

Finally, panelists will be given information about student performance on a test booklet. More specifically, panelists will be told the average percentage of items in test booklet that students at each achievement level would get correct. The test booklet will be the one administered to each panelists in the beginning of the process, so they will be rather familiar with the items in the booklet prior to this exercise. Again, this procedure is to be tested in the pilot studies and evaluated for incorporation in the ALS process.

■ ***Recommendations: Cutpoints, Descriptions and Exemplar Items***

Following the final round of ratings, panelists will be engaged in exercises aimed at providing information regarding their evaluations and perceptions of the achievement levels they have set. This step represents the **sixth element** in the process of setting achievement levels. Information will be collected from panelists regarding the percentages of students at their grade level expected to be at or above the cutpoint of each achievement level set in the final round. They will be presented with

information regarding the actual distribution of student scores, relative to the achievement levels. And, they will be asked to comment on the differences.

Information will also be collected from panelists about the score (over all items in their rating pool) they would estimate for students at each achievement level. Panelists will be presented with the empirically-based overall percentage of items that students at each achievement level would get correct, and they will be asked to comment on the differences between the estimated and the empirically-based differences.

Panelists will be asked to recommend an achievement level or distribution of student achievement and to comment on their rationale for that recommendation. Items in the 1994 assessment pool slated for public release by NAEP will be evaluated by panelists for possible recommendation as items to illustrate knowledge and skills associated with each achievement level. ACT will have classified items in the released pool according to statistical criteria. This statistical criteria will require a policy decision by NAGB to determine what level of student performance is necessary for consideration of an item as illustrative of an achievement level. (The items that would be flagged for consideration as illustrative of student performance at an achievement level would change considerably, depending upon the required level of probability of a correct response.)

Once that level has been set, however, the items can be identified for consideration by panelists. Panelists will be asked to recommend any items that seem appropriate for the purpose, given the criteria for consideration.

- ***Evaluation***

Evaluation is an integral part of the achievement levels-setting meetings, and this represents the **seventh element** in this process. At the end of each day, panelists will be asked to complete an evaluation form covering the activities of the day. These forms will be reviewed by facilitators each evening to identify any problems or concerns of panelists, and to determine whether the panelists are experiencing any problems with performing the tasks. The final evaluation form collects information for the process as a whole, and can be compared to evaluations at each intermediate step throughout the process. As nearly as practicable, the information collected will be the same as that collected from panelists during the 1992 ALS process, which is also quite similar to that collected during the 1990 process. This allows comparisons across rating sessions, content areas, and years to determine how panelists respond to different aspects of the achievement levels-setting process. The information collected from the panelists' evaluation forms is valuable in evaluating the impact of various feedback information and elements in the ALS process.

- ***Final Wrap-Up***

The **final element** of each ALS process is a meeting of the whole group following the final rounds of ratings and validation information sharing. This final element is an important element because it instills a final sense of closure to the process. At this

meeting of the whole, ACT proposes that the entire achievement levels-setting process be reviewed to show how each task and element contributed to the final result. The group will also be provided a review of the achievement level definitions, descriptions, and sample items for each grade level and a review of the recommended achievement levels that resulted from the analysis of the third round of ratings, as well as available information from the validation studies. Also at this final group meeting, evaluation instruments and comment forms will be administered and collected in order to provide ACT, NAGB, and the TACSS with evaluative feedback of the entire process.

Table 2

1994 ALS Procedures for Pilot Studies

Activity	All Panelists					
	Grade 4		Grade 8		Grade 12	
	Group A	Group B	Group A	Group B	Group A	Group B
Orientation	✓					
Frameworks	✓					
Revise Descriptions	✓		✓		✓	
NAEP Exam	✓	✓	✓	✓	✓	✓
Training 1	✓	✓	✓	✓	✓	✓
Practice Ratings	✓	✓	✓	✓	✓	✓
Evaluation 1	✓		✓		✓	
Round 1:						
Dichotomous Items	✓	✓	✓	✓	✓	✓
Polytomous Items[1] (geog.) (history)	P P	H H	M P[2]	P P	H H	M P[2]
Score Estimates	✓	✓	✓	✓	✓	✓
Performance Distribution	✓	✓	✓	✓	✓	✓
Feedback 1:						
Cutpoints & SDs	✓	✓	✓	✓	✓	✓
P-Values	✓	✓	✓	✓	✓	✓
Interjudge Consistency	✓	✓	✓	✓	✓	✓
Whole Booklet Exercise	✓	✓	✓	✓	✓	✓
Score Estimates	✓		✓		✓	
Adjust AL Descriptions	✓		✓		✓	
Training 2	✓	✓	✓	✓	✓	✓
Evaluation 2	✓	✓	✓	✓	✓	✓

Activity	All Panelists					
	Grade 4		Grade 8		Grade 12	
	Group A	Group B	Group A	Group B	Group A	Group B
Round 2:						
Dichotomous Items	✓	✓	✓	✓	✓	✓
Polytomous Items (geog.) (history)	P P	H H	M P[2]	P P	H H	M P[2]
Score Estimates	✓	✓	✓	✓	✓	✓
Performance Distributions	✓	✓	✓	✓	✓	✓
Feedback 2:						
Cutpoints & SDs	✓	✓	✓	✓	✓	✓
Interjudge Consistency	✓	✓	✓	✓	✓	✓
Intrajudge Consistency	✓	✓	✓	✓	✓	✓
Whole Booklet Scores	✓	✓	✓	✓	✓	✓
Can Do/Can't Do Exercise	✓	✓	✓	✓	✓	✓
Performance Distributions	✓		✓		✓	
Modify AL Descriptions	✓	✓	✓	✓	✓	✓
Training 3	✓	✓	✓	✓	✓	✓
Evaluation 3	✓	✓	✓	✓	✓	✓
Round 3:						
Dichotomous Items	✓	✓	✓	✓	✓	✓
Polytomous Items (geog.) (history)	P P	H H	M P[2]	P P	H H	M P[2]
Evaluation 4	✓	✓	✓	✓	✓	✓
Feedback 3:						
Cutpoints & SDs	✓	✓	✓	✓	✓	✓
Interjudge Consistency	✓	✓	✓	✓	✓	✓
Intrajudge Consistency	✓	✓	✓	✓	✓	✓
Whole Booklet Scores	✓	✓	✓	✓	✓	✓

Activity	All Panelists					
	Grade 4		Grade 8		Grade 12	
	Group A	Group B	Group A	Group B	Group A	Group B
Can Do/Can't Do Exercise	✓	✓	✓	✓	✓	✓
Score Estimates	✓	✓	✓	✓	✓	✓
Performance Distribution	✓	✓	✓	✓	✓	✓
Select Exemplar Items	✓		✓		✓	
Wrap-Up	✓					
Evaluation 5	✓					

[1] P = estimation of score point percentages; P^2 = estimation of percentage scored ≥ 2; M = estimation of mean scores; and H = the hybrid method.

Figure 3

"Draft" Agenda for Research Topics and Procedures to be Tested in the Pilot Studies of the 1994 Achievement Levels-Setting Process

DAY 1

Whole Group

2:30 P.M. Registration: name-tags and sign-in/information sheets.

3:00 P.M. Welcome and General Orientation Session
Introduction of Staff and Other Personnel; Explain How Selected; Review Agenda; Present Datashow with Flow Chart of Process

4:30 P.M. NAEP/NAGB Review: Evolution of NAEP; Role of NAGB; Policy on Achievement Levels; Purpose of Achievement Levels; Development of (Subject Area) Assessment

5:30 P.M. Get-Acquainted Social Time

6:00 P.M. Dinner

Grade Groups

7:30 P.M. Get Acquainted; Explain concept of item rating groups and table groups; Assign Table Groups; *Administer NAEP Exam for Grade Level* (two forms per grade); describe scoring guides and rubrics and have panelists self-score NAEP; "Assign" Review of Framework Booklet for Next Day

9:00 P.M. Evaluation of Day 1 Activities; Understanding of concepts, purposes, tasks, and so forth.

9:30 P.M. Adjourn

DAY 2

Whole Group

8:00 A.M. Continental Breakfast

8:30 A.M. Review Process *via* Datashow; Explain purpose for expanding and refining achievement levels descriptions; Provide parameters for changes to descriptions

Day 2
(Continued)

9:00 A.M. Presentation of Frameworks with extensive examples and illustrations to guide panelists; Present policy definitions and initial achievement levels descriptions; discuss process to operationalize preliminary descriptions.

10:15 A.M. Break

Grade Groups

10:30 A.M. Work toward reaching agreement on achievement levels descriptions. (Begin with mall group interaction to verbalize individual understanding of initial descriptions. Build to table groups, Item Rating Groups, and grade group. Begin setting goals for grade level with respect to task of expanding and refining descriptions.)

Whole Group

NOON LUNCH

Grade Groups

1:00 P.M. Continue work toward reaching agreement on achievement levels descriptions.
(Participate in focus exercises and practice applying descriptions using item sets.)

2:45 P.M. Break
(Facilitators and content consultants will meet to exchange descriptions for each grade level.)

3:00 P.M. Review descriptions of all grade groups. Continue working on grade-level descriptions; reach agreement on working versions.

Whole Group

4:00 P.M. Training for Rating Items. All panelists will learn about all item rating methodologies to be implemented during the pilot studies. In addition, all panelists will be trained in providing score estimates and %≥ estimates. Each panelists will have been told which rating methodology he/she will be using.

Day 2
(Continued)

Grade Groups

5:30 P.M. Practice item ratings.

Evaluation of Day 2 Activities
Evaluate procedures for arriving at agreement on achievement levels descriptions; evaluate training for rating items.

6:30 P.M. Adjourn

DAY 3

Whole Group

8:00 A.M. Continental Breakfast

Grade Groups

8:30 A.M. Review achievement levels descriptions; review training for rating items.

9:00 A.M. Round 1 Ratings

Whole Group

NOON LUNCH
(A buffet will be served to allow panelists to eat as they finish Round 1 ratings.)

3:00 P.M. Review Results of Round 1 Ratings
Training in Whole Booklet Exercise

Grade Groups

4:00 P.M. Whole Booklet Exercise
(NAEP test booklet administered on Day 1 will be used to present information on student performance for items included in the booklet as a whole.)
Evaluate Achievement Levels Descriptions and Make Agreed-upon Changes

Day 3
(Continued)

Grade Groups

6:00 P.M. Evaluation of Day 3 Activities
(Understanding of both rating tasks and feedback information; adequacy of training for rating and for using feedback information; confidence in ratings and ability to use feedback information; understanding of achievement levels descriptions and confidence in ability to use in rating; and so forth.)

6:30 P.M. Adjourn

DAY 4

Whole Group

8:00 A.M. Continental Breakfast

Grade Groups

8:30 A.M. Review Results of Round 1, Evaluation of Achievement Level Descriptions, and Feedback Data:
(Estimated P-values for Item Rating Pool and Interjudge Consistency Data Provided and Explained)
Training for:
- using feedback information (including score estimates);
- Round 2 ratings

10:00 A.M. BREAK

10:15 A.M. Treatment Group Feedback on Score Estimates

10:30 A.M. Round 2 Ratings

Evaluation of Round 2 Ratings: Training, understanding, confidence, and so forth.
(Evaluation of score estimate feedback information and effect on ratings: Treatment Group Panelists)

Whole Group

4:00 P.M. Review Results of Round 2 Ratings; Training for Item Mapping ("Can Do/Can't Do") Exercise

Day 4
(Continued)

Grade Groups

4:45 P.M. Review Results of Round 2 Ratings:

- Interjudge Consistency Data Updated and Presented for Each Item Rating Group (Control and Treatment)
- Whole Booklet Scores Updated
- Item Mapping Exercise Implemented
- Evaluate Achievement Levels Descriptions and Modify as Agreed

Note: Dinner will be provided for panelists in the grade group meeting rooms. We will try to arrange for something nutritious *and* easy to eat.

9:00 P.M. Evaluation of Item Mapping Exercise with respect to understanding of achievement levels descriptions and possible modifications of descriptions and with respect to sense of confidence in ratings using descriptions

DAY 5

Whole Group

8:00 A.M. Continental Breakfast

8:30 A.M.

- Review of rating methodology
- Review of achievement level descriptions for each grade
- Training in use of intrajudge consistency data
- Training in use of feedback information on %≥ estimates

Whole Group

9:30 A.M. BREAK

Grade Groups

9:45 A.M. Provide feedback on %≥ estimates to treatment group panelists

10:00 A.M. Distribute intrajudge consistency data
Review training and feedback information
Round 3 Ratings
Evaluation of Round 3 Ratings: Training, understanding of tasks, confidence in ability to perform tasks, confidence in ratings

DAY 5
(Continued)

Whole Group

NOON LUNCH

12:30 P.M. Wrap-Up/Summary of Activities to Date
(This will be the final Whole Group Session, and it will take place at the Lunch Site)

Grade Groups*

1:00 P.M.

- Review training in feedback to be receive and explanation of purpose(s) for collecting information
- Training in remaining tasks to perform (selection of illustrative items)
- Highlight evaluation feedback from panelists
- Evaluation of process from operations perspective
- Results of Round 3 Ratings
- Feedback on Interjudge Consistency (By Item Rater Group)
- Inform panelists about % ≥ at each achievement level; get feedback from panelists
- Inform panelists about Item Rating Pool Scores; get feedback from panelists
- Select Illustrative Items for Each Achievement Level
- Evaluation of Final Information; Evaluation of Entire Process

5:00 P.M. Adjourn

* Because ratings will have been completed by Grade 4 panelists first, the data analysis for their ratings, feedback, and statistical flagging of items for the selection of illustrative items can be completed earlier. In order to minimize the number of persons who will have to stay over another night before flying home, the remaining tasks will be conducted in Grade Groups. We anticipate that these activities will require approximately 2 - 2½ hours to complete. The Fourth Grade Group should be ready to adjourn by 3:30 P.M.. Processing and preparing data for the Eighth and Twelfth Grade groups will begin later, so the "turn around" may not be completed until 2:30 P.M. We feel that we must make plans to accommodate from one-third to one-half of the panelists on the evening of Day 5.

Section 4—Statistical Analyses

Four major goals have ben set for the statistical analyses to be performed in support of the Achievement Levels-Setting process for the 1994 NAEP.

- To support the rating process in both the pilot studies and the actual ALS process;
- To determine the effects of those experimental procedures for the pilot studies which are described in Section 3;
- To compare procedures for rating graded response items and determine which to use for the ALS process; and
- To gather information about the reliability and validity of the proposed ALS process.

Results from the pilot studies will help formulate the final design of the actual ALS process. This section of the design document indicates the analyses that are to be performed relative to each of these goals and the means by which the results of the analyses will be reported. Since the analyses necessary for the pilot studies will encompass the analyses needed for the actual ALS process, a more detailed description about the analyses for pilot studies will be presented first. The possible similarities and differences between the analyses for pilot studies and those for actual ALS will then be pointed out.

Analyses for the Pilot Studies

▪ ***Analyses to Support the Rating Process***

These analyses include summarizing the rating scores and transforming them into a NAEP-like theta scale using item parameters to calibrate a three-parameter IRT model. The theta values will be used to produce feedback data to promote intrajudge and interjudge consistency and to produce feedback data to inform pilot study panelists for the experimental procedures.

The rating methodologies have been described in the previous section. The ratings will be transformed onto the theta scale using the corresponding item characteristic curves. The result will be the theta values for each panelist for each item and for each achievement level.

The intrajudge consistency data are a distribution of theta values over different items for each panelist for each achievement level. Each panelist will receive three distributions (one for each achievement level) to examine the extent to which his/her ratings are consistent over items and what items correspond to outlier ratings. The distributions will be presented graphically, and an example is shown in Figure 4.

The interjudge consistency feedback data will be computed from the mean theta value of each panelist. A distribution of these mean theta values over panelists will be developed for the panelists to review. These distributions will be used to determine the level of agreement among panelists and to identify panelists whose ratings can be considered as outliers. Nine distributions will be computed in all, one for each achievement level at each grade level. An example is presented in Figure 5.

Four experimental procedures will be implemented in the pilot studies to test their utility for improving the outcome of the ALS process. These procedures are described in Section 3. They are: (a) Estimates of Performance Distributions; (b) Score Estimates; (c) Test Booklet Performance Data; and (d) Evaluations of Achievement Level Descriptions and Student Performance. The first two procedures (a and b) will be implemented with the experimental groups only, but the latter two procedures (c and d) will be implemented with all the panelists. Each of these procedures involves giving panelists feedback information that has the potential for influencing the ratings of panelists, their perceptions with respect to the descriptions of achievement levels, or both. The first two procedures, for example, involve presenting feedback information that seems more likely to influence panelists to focus on the numerical aspects of the rating process and, thus, to adjust their ratings. The latter procedure presents feedback information most likely to focus panelists' attention on the descriptions of the achievement levels. This might lead to new perceptions that would lead to adjustments in the achievement level descriptions, adjustments in achievement levels ratings, or adjustments in both.

Computations of feedback data for the Performance Distributions procedure will be based on the average of the equivalent theta values of each panelist's rating scores. Assuming that examinees' theta scores are normally distributed, the proportions of students at or above the cutpoints for each of the three achievement levels can be estimated for each panelist and averaged over the group. Each panelist in the experimental group will be informed about the percentages of the NAEP examinees at or above each of the three achievement levels based on his/her ratings. Percentages based on their average ratings will also be presented to that group.

In implementing the Score Estimates procedures, panelists will be asked to estimate the cutscores on each item block for each achievement level. The cutscores on all the blocks in the pool will be summed to get a total score for each panelist. Also, the item-by-item ratings will also be summed and compared to the block-by-block total scores. The comparisons will be presented to each panelist in the experimental group as feedback data to be evaluated.

The feedback information for the Booklet Performance Data procedure will be based on the achievement level cutpoints set by the panelists at each grade level and on the parameters for each item within the given test booklet. Based on the test characteristic curve for the booklet, the true score estimates can be found for all corresponding theta scores. The true scores are divided by the maximum possible score for that booklet to get average percentages correct for the booklet. The

estimated percentage correct for each achievement level will be presented to panelists as feedback information.

The implementation of the Evaluations of Achievement Level Descriptions and Student Performance procedure requires the classification of all the items into three categories given the achievement level cutscores at that point in the rating process, i.e., after Round 2. A detailed description of these categories can be found in the previous section. For each achievement level, three sets of items classified as described will be presented to the panelists as feedback information. This information will facilitate the panelists in examining whether there is correspondence between the knowledge and skills called for in the descriptions of each of the achievement levels and the knowledge and skills demonstrated by "student performance" at each of the achievement levels.

▪ ***Analyses to Determine the Effects of the Experimental Procedures***

Among the four experimental procedures proposed for the pilot studies, procedures (a) and (b) will be implemented only with the experimental group while procedures (c) and (d) will be implemented with all panelists at each grade level. The main purpose of the two experimental procedures is to give individualized feedback information so that panelists can use the information to adjust their rating if they wish. The first of the experimental procedures (a) will be implemented before Round 3 only and the second procedure (b) will be implemented before Round 2 only. Since both procedures are to be implemented with the same group, the possible effects will inevitably be confounded to a certain extent after both procedures are implemented. However, since they are implemented before different rounds, their effects can be quantitatively separated if certain assumptions can be met. The assumptions are these.

(1) Once a procedure is implemented, its effect will remain constant through the remaining rounds of the process.

(2) The effects are linearly additive, i.e., if the effect of one procedure can be quantified as, say, "3", and the effect of the other procedure can be quantified as "5", then after both procedures are implemented, their aggregate effect will be "8".

For analyzing the effects of procedures (a) and (b) the criterion variables are the three achievement level cutscores on the theta scale. The differences between the experimental group and the control group after Round 2 are composed of the effect of procedure (b) plus random error; those differences after Round 3 are composed of the aggregate effect of both procedures plus random error. The magnitude of random errors can be estimated based on the following sources of information: (1) the differences between the two group after Round 1; and, (2) the standard deviations within each group after each round.

The main purpose of procedures (c) and (d) is to provide panel-based feedback so that panelists can use this information to adjust their perception about the achievement levels and arrive at group agreement on changes in the descriptions that appear to be needed. Since they are implemented on both groups and their effects are manifested mainly through the change of the achievement level descriptions, their effects can only be analyzed qualitatively through the examination of the change in the achievement level descriptions, and through panelists' self evaluations of the effects.

▪ ***Analyses to Determine the Best Procedure for Rating Graded Response Items***
Three procedures for rating the extended response items will be tried out on an experimental basis in the pilot studies: (1) Estimated Mean Scores; (2) Estimated Score Point Percentages; and (3) The Hybrid Method. These procedures are described in detail in Section 3. Ratings by the panelists using one of the three procedures will be converted to mean score ratings for each graded response item and those mean score ratings will be transformed onto the theta scale using the corresponding test characteristic curve.

There is a lack of rigorous criteria to judge the merits of those procedures. Some plausible criteria can be postulated, however. They include the following:

1. Select the procedure that yields results closest to those yielded by ratings on the dichotomously scored items.

2. Select the procedure that is logistically the simplest and operationally the least confusing to the panelists.

3. Select the procedure that yield the most reliable results in terms of intrajudge and interjudge consistency.

Statistical analyses will be focused mainly on the first and last of these possible criteria. The second criterion will be evaluated qualitatively through during implementation and de-briefing sessions and through examining panelists' evaluations. If any procedure yields the best results on all three criteria, then it can be concluded that this procedure is *the* best procedure. Otherwise, more precaution needs to be taken to make the judgement.

Each of the three procedures will be implemented across two grades. This allows us to examine whether the relative "goodness" of the procedures can be generalized across grades.

Potential Limitations of Statistical Analyses for the Pilot Studies

There are two potential problems in performing the analyses for the pilot studies, as described above. One is that the item parameters available for analyses are based on field-trial data while the items used for ratings have been revised after the field

trials. There will be an inconsistency between items and item parameters. Since the item parameters are used through the entire process for providing various kinds of feedback information, the inaccuracy of item parameters poses a potential source of error.

The other problem is that in the pilot study, a number of experimental procedures will be tested and a rather complicated design is required to accomplish this. The formulation of the design is limited by the logistical possibilities of the studies and cannot be viewed as a rigorous research design. For instance, some procedures are implemented on the same group, and their effects will be confounded to some extent. To separate these effects requires that some rather strong assumptions be met.

The accuracy of these analyses will be affected by both these problems. As stated in Section 3, however, the general principle for the pilot studies is that collecting procedural information is more important than the accuracy of the numerical results. The results of quantitative analyses should shed light on the research questions proposed for the pilot studies and should be combined with various qualitative data to make the final judgement.

Analyses for the Actual ALS Process

The exact design for the actual ALS process will depend on the results from the pilot studies. Logically, one expects that the analyses to be performed for the actual ALS process will be a subset of those performed in the analyses for the pilot studies.

No experimental procedures will be implemented in the actual ALS process. The two groups in each of the three grades will use the same rating methods and receive the same information. This is obviously necessary in order to set a cutpoint across all items in the assessment at a particular grade level. The item rating pools are "matched" to be as equivalent as possible, the training provided to panelists is the same, and the rating methodologies are the same.

Figure 4

Intrarater Consistency Feedback

Figure 2.x. **A Consistent Panelist**

RATING CONSISTENCY SUMMARY PANELIST CODE: Y
ROUND NO. 2
ITEM FILE: HIST08A.ITM
NO. OF ITEMS = 25

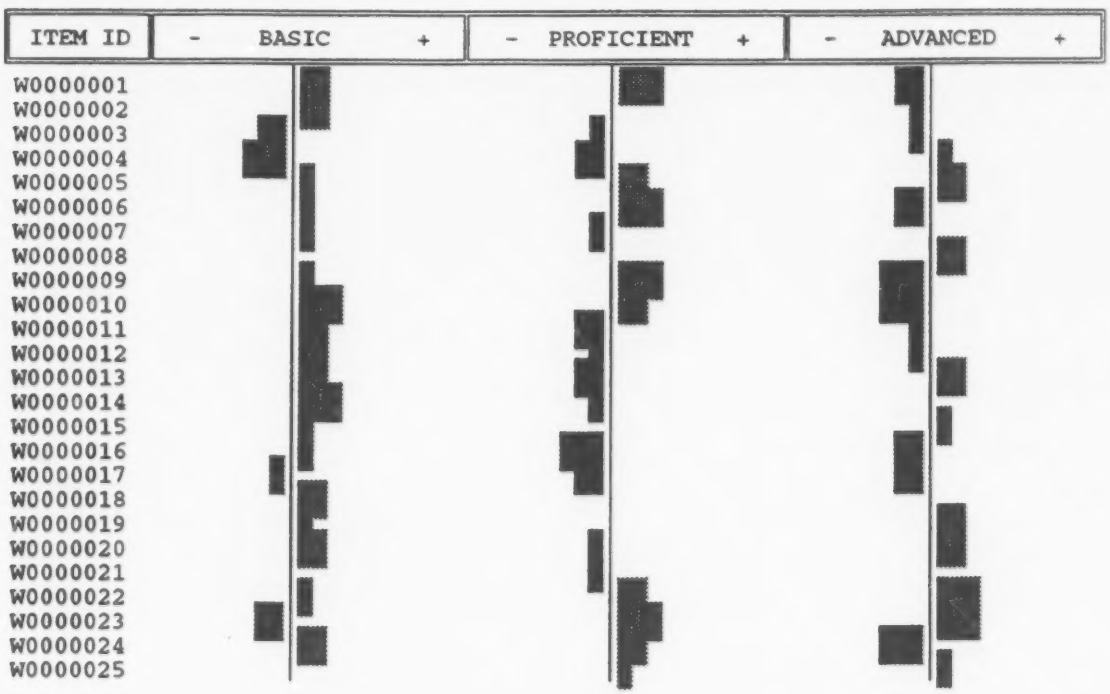
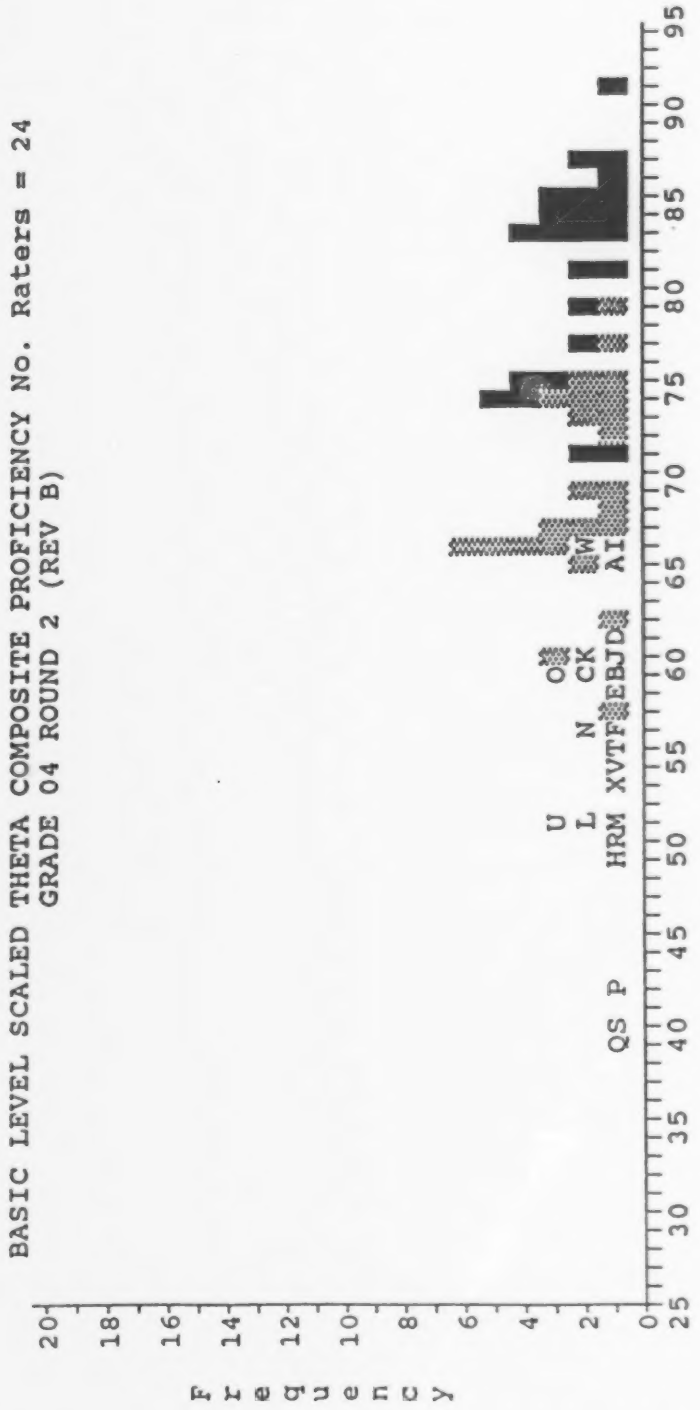


Figure 5

Interrater Consistency Feedback



Section 5—Public Comment Forums

The public comment solicitation will be conducted to encourage the same wide, thorough and open participation that ACT aims to have characterize this entire process. ACT proposes to hold two public comment forums for each subject area in order to increase the possibility that interested persons and groups can be represented. Moreover, ACT recommends that the meetings be held in different regions of the country.

On March 13 and March 15, the first public comment forum on each subject area will be held in Washington, D.C. The Washington, D.C. area was chosen because it is the headquarters site of most education organizations, as well as the location of many interested groups, such as Congressional Committee staffers, members of Congress, and the General Accounting Office.

On March 20 and 22, the second public comment forum on each subject area will be held in Denver, CO.

Several strategies of public notifications will be used to promote these meetings:

1) Announcements of the Public Comment Forums and invitations to attend shall be extended in various ways. First, notification of forum times, dates, and locations will be in the Federal Register. Second, this information will be mailed to individuals and groups represented at the achievement levels-setting meetings. Additionally, invitations will be mailed to those groups identified by NAGB and ACT as potentially interested in the ALS process.

2) ACT's Publications Department will develop promotional materials to publicize the forums. Color brochures will be developed and mailed to targeted audiences (e.g., principals, curriculum coordinators, social studies teachers, and school-level administrators; district superintendents, curriculum coordinators, and School Board presidents; mayors, commissioners, and leaders of Chambers of Commerce in local jurisdictions) in each metropolitan area at which forums will be held, as well as to targeted audiences at the state level (e.g., members of CCSSO and education committee members of state legislatures) for the host state and neighboring states (if geographically feasible).

3) Press packets will be developed and distributed to media outlets in the targeted areas. Television stations, radio stations, and newspapers will be targets. Announcements of the forums will be placed in widely accessible professional and popular media (e.g., The Chronicle of Higher Education, Education Week, USA Today). Finally, NAGB-authored press releases will be distributed through NAGB's usual information channels.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Bourque, M.L. (1993). The NAEP mathematics achievement levels. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Conaway, L. E. (1979). Setting standards in competency-based education: Some current practices and concerns. In M. A. Bunda & J. R. Sanders (Eds.), Practices and problems in competency-based education (pp. 72-88). Washington, DC: National Council for Measurement in Education.

Curry, L. (1987, April). Group decision process in setting cut off scores. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. Review of Educational Research, 59, 315-328.

Francis, A. S. & Holmes, S. E. (1980, August). Criterion-referenced standard setting in certification and licensure: Defining the minimally competent candidate. Paper presented at the annual meeting of the American Psychological Association, Anaheim, CA.

Friedman, C. B., & Ho, K. T. (1990, April). Interjudge consensus and intrajudge consistency: Is it possible to have both in standard setting? Paper presented at the annual meeting of the National Council on Measurement in Education, Boston, MA.

Geisinger, K. F. (1991). Using standard-setting data to establish cutoff scores. Educational Measurement: Issues and Practice, 10 (2), 17-22.

Huynh, H. (1985). Assessing mastery of basic skills through summative testing. In Daniel U. Levin, et al., Improving student achievement through mastery learning programs (pp.184-201). San Francisco, CA: Josey-Bass, Inc.

Johnson, E. G. & Mislevy, R. J. (1991, April). Theoretical background and philosophy of NAEP scaling procedures. In The technical report of NAEP's 1990 trial state assessment program. Washington, DC: National Center for Educational Statistics.

Kane, M.T. (1987). On the use of IRT models with judgmental standard setting procedures. Journal of Educational Measurement, 24, 333-345.

Kane, M.T. (1993). The validity of performance standards. Paper presented for the National Assessment Governing Board. Madison, WI: University of Wisconsin.

Klein, L. W. (1984, April). Practical considerations in the design of standard setting studies in health occupations. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores. Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Luecht, R.M. (1993, March). NAGB Round 4 Reading Results. Unpublished technical manuscript. Iowa City, IA: American College Testing.

Luecht, R. M. (1993, April). Using IRT to improve the standard setting process for dichotomous and polytomous items. Paper presented at the Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education, Atlanta, GA.

Luecht, R.M. (1993, April). Some results on the stability of the NAEP achievement level standards in mathematics. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.

Mills C. N., Melican, G. J., & Ahluwalia, T. (1991). Defining minimal competence. Educational Measurement: Issues and Practice, 10(2), 7-10.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. Applied Psychological Measurement, 14, 59-71.

National Academy of Education (1993). Setting Performance Standards for Student Achievement. National Center for Education Statistics, Washington,D.C.: U.S. Government Printing Office.

National Assessment Governing Board (1990). NAGB achievement levels policy (Appendix A in Request for Proposals). Washington, DC: Author.

National Assessment Governing Board (1993). Work statement for developing achievement levels on the 1994 National Assessment of Educational Progress in U.S. History and World Geography and Future Science Assessment (Attachment A in Request for Proposals). Washington, DC: Author.

National Council on Education Standards and Testing (1992). Raising standards for American education: A report to Congress, the Secretary of Education, the National Education Goals Panel and the American People. Washington, D.C.: U.S. Government Printing Office, ISBN 0-16-036087.

Plake, B. S., Melican, G. J., & Mills C. N. (1991). Factors influencing intrajudge consistency during standard-setting. Educational Measurement: Issues and Practice, 10(2), 15-16, 22, 25.

Reid, J. B. (1991). Training judges to generate standard-setting data. Educational Measurement: Issues and Practice, 10(2), 11-14.

Shepard, L. (1979). Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), Practices and problems in competency-based education (pp. 59-71). Washington, DC: National Council for Measurement in Education.

Shepard, L. (1980a). Standard setting issues and methods. Applied Psychological Measurement, 4, 447-467.

Shepard, L. (1980b). Technical issues in minimum competency testing. In D. C. Berliner (Ed.), Review of research in education, (pp. 30-84). Washington, DC: American Educational Research Association.

Shepard, L. (1983). Standards for placement and certification. In S. B. Anderson and J. S. Helmick (Eds.), On educational testing (pp.61-90). San Francisco, CA: Jossey-Bass.

Smith, R. L. & Smith, J. K. (1988). Differential use of item information by judges using the Angoff and Nedelsky procedures. Journal of Educational Measurement, 25, 259-274.

Appendix A

Organizations to be Contacted for Input Related to Selection of Panelists, and Related Materials

Listing of Organizations to be Contacted

Alliance for Educ. in Global and International Studies
American Association for Continuing Education
American Association for Parents and Teachers
American Association for the Advancement of Science
American Association of School Administrators
American Association of University Professors
American Business Women's Association
American Classical League
American Council for the Arts
American Council of Learned Societies*
American Council of Teachers of Foreign Language
American Educational Research Association
Amer. Federation of Labor-Congress of Indust. Orgns.
American Federation of Teachers
American Geographical Society
American Historical Association
American Indian Heritage Foundation
American Legislative Exchange Council
American Public Transit Association
American Society of Transportation and Logistics
American Society of Travel Agents
Associated Motor Carriers Tariff Bureau
Association for Supervision and Curriculum Development
Association of American Geographers
Association of Boarding Schools
Association of Community Travel Clubs
Association of Corporate Travel Executives
Association of Retail Travel Agents
Association of Travel Marketing Executives
Business Roundtable
Carnegie Foundation
CHART*
Coalition for Essential Schools
Commission for Economic Development
Council for American Private Education
Council for Basic Education
Council of Chief State School Officers*
Council of Great City Schools
Delaware Department of Public Instruction
Dwight D. Eisenhower Library
Education Commission of the States
Educational Information & Advisory Committee
Educational Testing Service
Evaluation Assistance Center, East
Franklin D. Roosevelt Library
Friends Council on Education
Friends for Education
General Counsel for Education
Gerald R. Ford Library*
Harry S. Truman Library*
Herbert Hoover Library*
Hispanic Policy Development Project
Illinois State Board of Education
Jimmy Carter Library*
John F. Kennedy Library
Joint Council on Economics Education
Latin American Educational Foundation
League of United Latin American Citizens
Lutheran Education Association
Lyndon B. Johnson Library
Mexican-American Legal Defense and Educational Fund
National Academy of Sciences
National Alliance of Business
National Art Education Association
National Assoc. for Asian and Pacific American Educ.
National Association for Industry-Education Cooperation
National Assoc. for the Advancement of Colored People
National Association of Elementary School Principals
National Association of Episcopal Schools
National Association of Female Executives
National Association of Manufacturers
National Association of Professional Educators
National Association of Secondary School Principals
National Association of State Boards of Education
National Association of Test Directors
National Catholic Education Association
National Center for Fair & Open Testing
National Conference of State Legislatures
National Congress of Parents & Teachers
National Council for Geographic Education*
National Council for History Education*
National Council for Social Studies
National Education Association
National Federation of Independent Business
National Geographic Society
National Governors' Association
National Indian Education Association
National Middle Schools Association
National Minority Business Council
National Park Service
National School Boards Association*
National Science Teachers Association
National Urban League
Nixon Presidential Materials Staff
Office of Congressman Matthew Martinez
Office of Senator Jeff Bingaman
Organization of American Historians
Quality Education for Minorities Network
Ronald Reagan Library
Society of History Education
State Departments of Education (KY)*
(50 states plus the District of Columbia and 5 territories)
Teacher Networks Group Project
The Asia Society
The Aspen Institute
The College Board
The Heritage Foundation
U.S. Chamber of Commerce
U.S. General Accounting Office
U.S. House of Representatives
U.S. Senate

* Provided written comments

81 BEST COPY AVAILABLE

Stakeholder Meetings to Review Draft ALS Design Document
Suite 370, One Dupont Circle, Washington, D.C.
February 22-25, 1994

Letters were sent to invite participation of approximately 200 groups, organizations, and individuals identified as potential stakeholders for the U.S. History NAEP and Geography NAEP Achievement Levels. The first letters were sent on January 26 to notify the stakeholders that they would soon receive a draft version of the Design Document. That mailing included a form for signing-up to meet in Washington or for agreeing to send written comments. The Design Document draft was mailed to these stakeholders on February 8.

A total of 19 persons signed up for meetings in Washington, and 15 attended. (Attendees are listed below.) The meetings provided a much-needed opportunity to meet face-to-face with representative of key organizations and individuals who have a high level of interest in the outcome of this project. Comments were generally very positive and supportive. The following are concerns and suggestions for changes.

- Comments indicated that readers had difficulty understanding the document—too much jargon. ACT suggested that a brief document, similar to an executive summary with key points highlighted, could be prepared for purposes of informing the stakeholders about the design and plans for carrying it out. This suggestion met with strong support. Several participants indicated that they would still prefer to receive the more technical and precise Design Document.

- Additional associations to include among stakeholders were made.

- Guidelines for nominators should include the need for panelists to be strongly grounded in skills as well as content.

- Practioners (teachers) are not aware of the current trends in teaching and assessing geography and history. If the persons chosen to serve as panelists do not accept the ideology/philosophy of the discipline represented in the framework document (for history or geography), then the panelists cannot reach agreement on the achievement levels to be set. How can you ensure that all teachers "buy into" the frameworks and the approach to the discipline embodied therein?

- Specify that the state curriculum director/coordinator to nominate teachers be the state *social studies* director/coordinator. (Note: Most states do not have a social studies curriculum coordinator listed among state school officials.)

- Students at the undergraduate college level would be particularly good general public panelists. (Some discussion about the choice of graduate students instead of undergraduate students. Problems identified with graduate students as representatives of "general public" since they are likely training to be "educators." Compromise suggestion was to have students serve in focus group activities to comment on outcomes of ALS process.)

- No radical exploration included in the design. Significant portions of the psychometric community will be involved in other standard setting methods. NAGB must weigh in with those. There should be "development" (as opposed to "research") included in the design. Nothing recommended by the NAE has been included in the design. Why not? The psychometric community will come in with the NAE on questioning whether the Angoff process is really a rating task that panelists can do.

The assessments are not a single continuum. The move by NAEP is toward more and more performance assessment tasks. These performance assessment tasks are not unidimensional, and the standard setting methodology cannot be unidimensional. Notwithstanding the fact that NAEP scaling is based on assumption of unidimensionality.) Achievement levels (and NAEP scores) should be reported for each purpose for reading, for example: one for each reading situation; not a single score/level.

- Concern that many classroom teachers will not be familiar with current aspects of the discipline incorporated in the frameworks. Suggested that some effort be made to ascertain whether panelists are "current" with respect to content, skills, and methodology of the discipline.

- Concerned with lack of expertise among panelists. Suggest a questionnaire to make certain nominees meet standards for being panelists. (Planned in design.)

- Concern with the nominators. Believe they will not be representative of the community specifically, that they will exclude language minorities. Teachers of language minority students will not be nominated because will not be perceived as "outstanding" by these nominators.

  Ask about bilingualism on the questionnaire regarding panelists' credentials and qualifications. Be specific about language diversity and cultural minorities.

Stakeholder Participant and Organization Represented

1. Osa Brand, **Association of American Geographers**
2. Frederick H. Brigham, Jr., **National Catholic Education Association**
3. Robert E. Dulli, **National Geographic Society**
4. Noralee Frankel, **American Historical Association**
5. James Goodman (Geographer-in-Residence) **National Geographic Society**
6. Louis Harlan, **Organization of American Historians**
7. Laurel Kanthak, **National Association of Secondary School Principals**
8. Barbara Kapinus, **Council of Chief State School Officers**
9. Denise McKeon, **American Educational Research Association**
10. Susan Munroe, **S.S. Munroe, Inc. (Geography Consensus Project Coordinator)**
11. Salvatore J. Natoli, **National Council for Geography Education**
12. Lois Osmer, **Maryland Department of Education**
13. Doris Redfield, **Virginia Department of Education**
14. Charlene Rivera, **Evaluation Assistance Center East**
15. Ramsey Selden, **Council of Chief State School Officers**

Susan Loomis, ACT Project Director, and Mary Lyn Bourque, NAGB Assistant Director for Psychometrics were present at all sessions to receive comments and discuss the design plans with participants. Mary Crovo (NAGB Staff) also attended one session.

Written Responses from Stakeholders

Approximately 200 groups, organizations, and individuals were sent a draft copy of the Design Document. Responses were received from approximately 50% of those contacted: 40 agreed to send written comments, but only 9 have responded to date. This response includes 9 who indicated that they were not interested in the project and would not be interested in receiving additional materials. Quite a few persons (27) responded that they were interested in the project, but were unable to attend the meetings or to provide written comments within the required time.

In general, the written comments were very thoughtful and will be very helpful. They indicated a bit of confusion on the part of the readers with respect to the purpose of the design document. This confusion had been discussed during the meetings described above. Additional comments included the following.

Appendix B

Nominator Materials
Sample Letters to Nominators and Nomination Forms

Guidelines for Selecting Teacher Nominees and Information for Potential Nominees

We ask that you nominate only those teachers whom you deem to be "outstanding." Outstanding teachers are those who are held in high regard by administrators, students, and/or fellow teachers, or who have been honored/recognized in some way, such as being named "teacher of the year." Teachers who have been very active in content-related professional associations, such as NCGE or the Geography Alliance, would be especially appropriate nominees.

We urge you to consult with your colleagues, with contacts you might have with local, state, or national content-related professional associations, or with others who might assist you in identifying your best geography teachers. We have also given your name to representatives of interested professional associations so that they might suggest names of teachers for you to consider.

Teacher nominees should have at least five (5) years of classroom experience, and must currently be classroom teachers. Two years of that experience (preferably most recent) should be in teaching students at the 4th, 8th, or 12th grade levels, and should, at a minimum, be in teaching courses in the social sciences with a geography component. Please nominate **up to** four (4) teachers for each grade level (4th, 8th, and 12th). The teachers that you nominate for a grade level must be teachers of students at that grade level. The teachers must teach geography courses, *per se*, or social science courses with a geography component (e.g., global studies, environmental studies, and Asian studies).

The 1994 NAEP Geography Achievement Levels-Setting process is scheduled for five-days, November 12-16, 1994. The meetings are scheduled to include a week-end in order to minimize the number of days teachers will need to be away from their classrooms. The meetings will be held at the Ritz-Carlton Hotel in St. Louis, Missouri, and all panelists will stay at the Ritz-Carlton throughout the meeting period. Panelists will not receive compensation for their participation *per se*, but their expenses—travel, lodging, and meals—will be paid. In addition, we will reimburse the school for the cost of hiring a substitute teacher for the days the teacher(s) selected are away.

We ask that you discuss this with the teachers that you might wish to nominate before you submit their names. We will appreciate your ascertaining that any teacher selected to serve as a panelist will have permission to do so and to be away from their normal teaching responsibilities during the meeting period. It is important for you and any nominees to understand, however, that this is a nomination only. **Not all teachers nominated will be selected as panelists.** We are requesting teacher nominations from four groups: superintendents, teacher association leaders, private school principals, and state curriculum supervisors. Moreover, we are requesting nominations of panelists in different categories (representing the general public and non-teaching educators) from other individuals throughout the nation, and we intend to select the most outstanding nominees. In addition, the final selection of teachers will be made in a way that will ensure that the panels are balanced with respect to gender, race/ethnicity, region of the nation, and other important characteristics. For that reason, we are asking you to identify the gender and race/ethnicity of teachers you nominate (see Nomination Form).

A pre-addressed, postage-paid envelope is enclosed for you to return your nominations. Please return your nominations at your earliest convenience, but please try to do so **by September 7, 1994**. In case you would prefer to FAX nominations, our FAX number is 319/339-3020.

Guidelines for Selecting Nonteacher Educator Nominees and Information for Potential Nominees

Nominees must be educators (K-12, college/university, district/state level personnel) who are **not currently** classroom (K-12) teachers. Nominees, for example, could include guidance counselors, curriculum specialists, assessment specialists, principals, former teachers who are now administrators, college faculty members, college admissions officers, teachers of college freshmen, educational researchers, and so forth. Nominees should be knowledgeable about the learning and skills levels of students at the 4th, 8th, and 12th grade levels; awareness and understanding of geographic knowledge skills is particularly relevant.

We urge you to consult with your colleagues, with contacts you might have with local, state, or national professional associations (e.g., NCGE, AAG, the Geography Alliance), or with others who might assist you in identifying outstanding non-teacher educators. We have also given your name to representatives of interested professional associations so that they might suggest names for you to consider.

Please nominate **up to** four (4) non-teacher educators for each grade level (4th, 8th, and 12th). **We encourage you to nominate yourself, if you so desire, and if you meet the specified criteria.**

The 1994 NAEP Geography Achievement Levels-Setting process is scheduled to last five days, November 12-16, 1994. The meetings are scheduled over a week-end in order to minimize the number of days panelists will need to be away from work. The meetings will be held at the Ritz-Carlton Hotel in St. Louis, Missouri, and all panelists will stay at the Ritz-Carlton throughout the meetings. Panelists will not receive compensation for their participation *per se*, but their expenses—travel, lodging, and meals—will be paid.

We ask that you discuss this with the individuals you might wish to nominate before you submit their names. If necessary, we can contact the supervisors of nominees who are selected as panelists to secure permission for panelists to participate and to be away from their normal work responsibilities during the meeting period. It is important for you and other nominees to understand that this is a nomination only. **Not all nominees will be selected as panelists**. We are requesting nominations from other individuals throughout the nation, and we intend to select the most outstanding nominees. In addition, the final selection of panelists will be made in a way that will ensure that the panels are balanced with respect to gender, race/ethnicity, region of the nation, and other important characteristics. For that reason, we are asking you to identify the gender and race/ethnicity of nominees, including yourself (see Nomination Form).

A pre-addressed, postage-paid envelope is enclosed for you to return your nominations. Please return your nominations at your earliest convenience, but please try to do so **by September 7, 1994**. In case you would prefer to FAX nominations, our FAX number is 319/339-3020.

Guidelines for Selecting Community Member Nominees and Information for Potential Nominees

Nominees **cannot** be **current** teachers or educators and, to the extent that you can determine, should not be former teachers or educators (K-12 or college/university). Nominees, however, **can** be members of local school boards (if they are not also employed as educators). Nominees should be knowledgeable about the learning and skills of students at the 4th, 8th, and 12th grade levels; awareness and understanding of geographic knowledge and skills would be particularly desirable. Nominees might also include people who utilize geographic skills extensively in their work or everyday lives. **Examples** of the types of people you might consider nominating include:

- active PTA/PTO members
- parents of elementary or secondary schoolchildren
- employees of transportation and shipping companies (e.g., dispatchers)
- former Peace Corps volunteers
- urban planners
- forest service employees
- National Parks Rangers
- travel agents
- marketing/locational analysts
- school board members
- business leaders with an interest in education
- personnel directors at local or regional businesses and industries
- members of local business groups, such as Business Round table or National Alliance of Business, who are actively interested in education
- local labor organizations

We urge you to consult with your colleagues, with members of local organizations that are actively interested in education, or with other contacts who might assist you in identifying individuals to serve as panelists. We have also given <u>your</u> name to representatives of interested professional associations so that they might suggest names to you.

Please nominate **up to** four (4) individuals for each grade level (4th, 8th, and 12th). **We encourage you to nominate yourself, if you so desire, and if you meet the specified criteria.**

The Achievement Levels-Setting process is scheduled to last five days, November 12-16, 1994. The meetings are scheduled over a week-end in order to minimize the number of days panelists will need to be away from work (if they work). The meetings will be held at the Ritz-Carlton Hotel in St. Louis, Missouri, and all panelists will stay at the Ritz-Carlton throughout the meetings. Panelists will not receive compensation for their participation *per se*, but their expenses—travel, lodging, and meals—will be paid.

We ask that you discuss this with the individuals you wish to nominate before you submit their names. If necessary, we can contact the supervisors of nominees who are selected as panelists to secure permission (if appropriate) for panelists to participate and to be away from their normal work responsibilities during the meeting period. It is important for you and the

(over)

nominees to understand that this is a nomination only. **Not all nominees will be selected as panelists**. We are requesting nominations from other individuals throughout the nation, and we intend to select the most outstanding nominees. In, addition, the final selection of panelists will be made in a way that will ensure that the panels are balanced with respect to gender, race/ethnicity, region of the country, and other important characteristics. For that reason, we are asking you to identify the gender and race/ethnicity of nominees, including yourself (see the Nomination Form).

A pre-addressed, postage-paid envelope is enclosed for you to return your nominations. Please return your nominations at your earliest convenience, but please try to do so **by September 7, 1994**. If you would prefer to FAX nominations, our FAX number is 319/339-3020.

A Brief Summary of Panelists' Responsibilities

Approximately two weeks prior to the meeting, all panelists will receive a packet that will include background and training materials for their review prior to the meeting. There will be approximately 30 panelists for each grade-level (4th, 8th, and 12th).

As part of the orientation to the process, panelists will receive an overview of the National Assessment of Educational Progress (NAEP), an explanation of the content area framework with which they will be working, and other training related to the NAEP. They will then study and evaluate definitions of Basic, Proficient, and Advanced performance in this content area and for their particular grade level and, expand and refine these definitions. They will be assisted in this task by content specialists who have worked with the development of the content frameworks and facilitators at each grade level who are trained in standard setting, assessment issues and the content frameworks. Each panelist will complete one form of their grade-level NAEP Assessment to familiarize them with the content and format of the test.

Panelists will discuss and modify their operational definitions of Basic, Proficient, and Advanced, reach agreement on common definitions for their grade-level, and receive copies of these definitions for later use. Panelists will then receive extensive training in the rating task for the achievement levels setting process (see below), including practice in carrying out that task.

The Task. Panelists' primary responsibility for setting achievement levels will be to examine individual items (test questions) for their grade-level NAEP. The panelists will be determining, for example, how 12th graders perfectly characterized by the achievement levels definitions will perform on each item in the NAEP pool for a particular grade level. With the definitions of Basic, Proficient, and Advanced student performance in mind, panelists will decide what percentage of 12th grade students who are Basic, will get that item correct, what percentage of 12th grade students who are Proficient will get that item correct, and what percentage of 12th grade students who are Advanced will get that item correct. Panelists will repeat this procedure for a specified number of test items (the number varies by content area and grade-level). We anticipate one round of such ratings will take an average of 3-4 hours. There will be three rounds of ratings. At the end of the third round of ratings, panelists' percentage correct estimates for each achievement level (Basic, Proficient, and Advanced) for each item will be used to compute an average percentage correct figure for Basic, Proficient, and Advanced, for the item pool. Each panelists' percentage correct estimate for Basic, Proficient, and Advanced will then be combined with the estimates of the other panelists at that grade level and averaged to produce a grade level estimate for each achievement level. The group estimate will represent the achievement level estimate for Basic, Proficient, and Advanced.

Panelists will receive retraining in the rating procedure before each round of ratings, information about the actual student performance on the items they rated, and information about the consistency of their ratings. We will engage panelists in several exercises and tasks designed to inform them about student performance. They will have the opportunity to review the achievement level descriptions and make additional refinements in light of the additional information and insights gained through rating items in the first two rounds. Agreement on the descriptions must be solidified for panelists at each grade level before the final round of ratings, however.

After the ratings have been completed, panelists will be given a complete review of the process and information gathered throughout the process will be shared to give a comprehensive view of the process, step by step. We will discuss the entire process with panelists, and show panelists how their individual contributions helped produce a final result.

Following each procedure and task in the process, we will have panelists evaluate their experiences. These evaluations will be reviewed carefully and used in the documentation of the process and success of the procedures.

The National Assessment of Educational Progress (NAEP) and the Achievement Levels-Setting (ALS) Process

The NAEP

The National Assessment of Educational Progress (NAEP), an official U.S. Department of Education program, has provided information on the achievement and performance of students in the U.S. for over two decades. For each assessment, a nationally representative sample of approximately 35,000 to 100,000 students drawn from three age or grade levels has taken tests in various subject areas. The resulting data on student knowledge and performance have been accompanied by descriptive information allowing analyses of a variety of student experiences and background factors that correlate with student achievement.

The assessments have been designed to allow comparisons of student performance over time and among sets of students, grouped by region, type of community, race/ethnicity, and gender. The NAEP, commonly referred to as "The Nation's Report Card," is the most comprehensive and **only** continuing, valid source of information on what U.S. students know and can do, and of how their performance has varied over time.

Achievement Levels-Setting

Public Law 100-297 (1988) contained the National Assessment of Educational Progress Improvement Act. The NAEP Improvement Act created the National Assessment Governing Board (NAGB) to formulate policy guidelines for NAEP and provided that NAGB's responsibilities include:

- "Taking appropriate actions to improve the form and use of the National Assessment; and
- Identifying appropriate achievement goals for each age and grade in each subject area to be tested under the National Assessment."

By defining levels of appropriate achievement on the National Assessment, NAGB seeks to increase greatly the significance and usefulness of NAEP results to educators, policymakers, and the American public. Moreover, it is consistent with the Clinton Administration's *Goals 2000: Educate America Act* to "develop voluntary academic standards and assessments that are meaningful, challenging and appropriate for all students. . .."[5]

To carry out these responsibilities as specified in the NAEP law regarding appropriate achievement goals, NAGB released a Request for Proposal on June 18, 1993 for setting achievement levels on the 1994 NAEP in U.S. History, Geography, and future Science assessment. After reviewing several proposals, a contract was awarded to American College Testing (ACT) to design and administer a process that would allow NAGB to establish achievement levels on the NAEP to specify what students **should** know and be able to do. These levels will be determined in accordance with the policy framework, definitions, and technical procedures in the NAGB policy titled, Setting Appropriate Achievement Levels for the NAEP, dated May 10, 1990.

The policy calls for three achievement levels with clear distinctions between them. The achievement levels will be established for each grade and subject tested under NAEP. These levels will be called:

---

[5] National Council on Education Standards and Testing, *Raising Standards for American Education: A Report to Congress, the Secretary of Education, the National Education Goals Panel, and the American People*. (Washington, D.C.: U.S. Government Printing Office, January 24, 1992, ISBN 0-16-036087).

Proficient: This central level represents solid academic performance for each grade level tested.

Basic: This level, below proficient, denotes partial mastery of knowledge and skills that are fundamental for proficient work at each grade level tested.

Advanced: This higher level signifies superior performance beyond proficient grade-level mastery.

It is NAGB's intention to use these three achievement levels as the primary means of reporting results for all newly developed assessments beginning with 1992. The system is in contrast to NAEP's past practice of simply describing how students perform with no reference to standards of how well they ought to do.

The process of determining achievement levels is to be a logical continuation of the national consensus effort used in developing the content and objectives of the NAEP. A broadly representative group of panelists will assist in defining the achievement levels using a proven judgment procedure to recommend levels of basic, proficient, and advanced in terms of the NAEP subject areas.

As part of their deliberations, the panelists will prepare detailed descriptions of the subject-matter knowledge and skills proposed for each achievement level. These descriptions will be illustrated by representative sample items and scoring protocols. In preparing descriptions of achievement levels and assigning test items to them, panelists will use their best judgment and expertise and will also take into account a wide range of background information and frames of reference provided by ACT.

If you, or potential nominees, desire additional information about the NAEP or the ALS process, feel free to contact Dr. Susan Cooper Loomis (ACT Project Director, 319/337-1048) or Ms. Luz Bay (ACT Assistant Project Director, 319/337-1639).

Date

drmrms~ firstname~ lastname~
title~
organization~
address~
city~, state~ zip~

Dear drmrms~ lastname~:

The National Assessment Governing Board (NAGB) invites you to nominate outstanding teachers from your district to serve as panelists for the 1994 National Assessment of Educational Progress (NAEP) Geography Achievement Levels-Setting Project. The study is scheduled to be held November 12-16, 1994, in St. Louis, Missouri, at the Ritz-Carlton Hotel. Enclosed are 1) a brief description of the NAEP program and the Achievement Levels-Setting process, 2) guidelines for selecting nominees, and 3) forms for listing nominees.

We would like to emphasize that you and your district, through the people you nominate, will be involved in helping set **national** standards in Geography for our nation's schoolchildren. In addition, those nominees who are selected as panelists will have a very positive developmental experience through learning about the NAEP testing program, the national consensus process that led to national content specifications for NAEP Geography Assessment, and will be able to interact and develop networks with other outstanding teachers from throughout the nation.

As an integral part of monitoring the national education goals, the setting of achievement levels for our nation's youth in Geography, and other academic content areas, is vitally important to the continued improvement of our nation's educational system. For that reason, nominators, such as yourself, were selected with great care. Nominations of teachers include district superintendents, principals (or equivalent) of private schools, and the leaders of the largest/bargaining representative teachers organization in sampled districts. Your participation in the nomination process is very important. A self-addressed, stamped envelope is enclosed for your reply. Or, if you prefer, you may FAX nominations to us at 319/339-3020.

Please call me (319/337-1048) or the Assistant Director, Luz Bay (319/337-1639) if you have **any** questions about the project, or about nominating individuals.

Thank you for your consideration and assistance.

Sincerely,

Susan Cooper Loomis, Ph.D.
Director
NAEP Project
Research Division

SCL:tjf

GRADE 4
Teachers

a. Name _____

Home Address _____

City/State/Zip _____

Home Phone ( ) _____

School Name _____

School Address _____

City/State/Zip _____

School Phone ( ) _____

b. Total Years Teaching Experience

- ☐ 5-9
- ☐ 10-14
- ☐ 15 or more

c. Total Years Teaching Subject

- ☐ 2-4
- ☐ 5-9
- ☐ 10 or more

d. Race/Ethnicity

- ☐ White
- ☐ Black
- ☐ Asian
- ☐ Native American
- ☐ Hispanic
- ☐ Other _____

e. Gender

- ☐ Male
- ☐ Female

f. Why do you feel this person is an outstanding candidate? (Please write on back or attach another page if more space is needed.)

_____

_____

_____

_____

_____

_____

Date

drmrms~ firstname~ lastname~
title~
organization~
address~
city~, state~ zip~

Dear drmrms~ lastname~:

The National Assessment Governing Board (NAGB) invites you to nominate outstanding nonteacher educators to serve as panelists for the 1994 National Assessment of Educational Progress (NAEP) Geography Achievement Levels-Setting Project. The study is scheduled to be held November 12-16, 1994, in St. Louis, Missouri, at the Ritz-Carlton Hotel. Enclosed are 1) a brief description of the NAEP program and the Achievement Levels-Setting process, 2) guidelines for selecting nominees, and 3) forms for listing nominees. Please note that you may qualify as a nominee too.

We would like to emphasize that you and your community, through the people you nominate, will be involved in helping set **national** standards in Geography for our nation's schoolchildren. In addition, those nominees who are selected as panelists will have a very positive developmental experience through learning about the NAEP testing program, the national consensus process that led to national content specifications for NAEP Geography Assessment, and will be able to interact with other outstanding educators from throughout the nation.

As an integral part of monitoring the national education goals, the setting of achievement levels for our nation's youth in Geography, and other academic content areas, is vitally important to the continued improvement of our nation's educational system. For that reason, nominators, such as yourself, were selected with great care. Nominators for this type of panelist include educators who are not K-12 classroom teachers from sampled school districts, state curriculum directors, and college/university academic leaders. Your participation in the nomination process is very important. A self-addressed, stamped envelope is enclosed for your reply. Or, if you prefer, you may FAX nominations to us at (319) 339-3020.

Please call me (319/337-1048) or the Assistant Director, Luz Bay (319/337-1639) if you have **any** questions about the project, or about nominating individuals.

Thank you for your consideration and assistance.

Sincerely,

Susan Cooper Loomis, Ph.D.
Director
NAEP Project
Research Division

SCL:tjf

GRADE 4
Nonteacher Educators

a. Name _____

Home Address _____

City/State/Zip _____

Home Phone ( ) _____

Employer/Company Name _____

Work Address _____

City/State/Zip _____

School Phone ( ) _____

b. Race/Ethnicity

- ☐ White
- ☐ Black
- ☐ Asian
- ☐ Native American
- ☐ Hispanic
- ☐ Other _____

c. Gender

- ☐ Male
- ☐ Female

d. How is this person familiar with the subject matter and/or content area of Grade 4 geography? Why would this person be an "outstanding" panelists? Please use as much space as needed to provide this information.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

Date

drmrms- firstname- lastname-
title-
organization-
address-
city-, state- zip-

Dear drmrms- lastname-:

The National Assessment Governing Board (NAGB) invites you to nominate outstanding members of your community to serve as panelists for the 1994 National Assessment of Educational Progress (NAEP) Geography Achievement Levels-Setting Project. The study is scheduled to be held November 12-16, 1994, in St. Louis, Missouri, at the Ritz-Carlton Hotel. Enclosed are 1) a brief description of the NAEP program and the Achievement Levels-Setting process, 2) guidelines for selecting nominees, and 3) forms for listing nominees.

We would like to emphasize that you and your community, through the people you nominate, will be involved in helping set **national** standards in Geography for our nation's schoolchildren. In addition, those nominees who are selected as panelists will have a very positive developmental experience through learning about the NAEP testing program, the national consensus process that led to national content specifications for NAEP Geography Assessment, and will be able to interact with other outstanding citizens from throughout the nation.

As an integral part of monitoring the national education goals, the setting of achievement levels for our nation's youth in Geography, and other academic content areas, is vitally important to the continued improvement of our nation's educational system. For that reason, nominators, such as yourself, were selected with great care. Nominators of the general public panelists include mayors, school board presidents, and chairs of education committees of local chambers of commerce in districts drawn from a national sample. Our goal is to have broadly representative panelists who are qualified to perform the tasks. Your participation in the nomination process is very important. A self-addressed, stamped envelop is enclosed for your reply. Or, if you prefer, you may FAX nominations to us at (319) 339-3020.

Please call me (319/337-1048) or the Assistant Director, Luz Bay (319/337-1639) if you have **any** questions about the project or about nominating individuals.

Thank you for your consideration and assistance.

Sincerely,

Susan Cooper Loomis, Ph.D.
Director
NAEP Project
Research Division

SCL:tjf

GRADE 4
Nominees to Represent the General Public

a. Name _____

Home Address _____

City/State/Zip _____

Home Phone ( ) _____

Employer/Company Name _____

Work Address _____

City/State/Zip _____

School Phone ( ) _____

b. Race/Ethnicity

- ☐ White
- ☐ Black
- ☐ Asian
- ☐ Native American
- ☐ Hispanic
- ☐ Other _____

c. Gender

- ☐ Male
- ☐ Female

d. How is this person familiar with the subject matter and/or content area of Grade 4 geography? Why would this person be an "outstanding" panelists? Please use as much space as needed to provide this information.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____